BRAZIL

RESOLUTION INPI/PR NO. 187, OF APRIL 27, 2017

Procedures for the Submission of the Sequence Listing in electronic form, in addition to the rules for the representation of the sequences of nucleotides and/or amino acids in the Sequence Listing.

Table of Contents

BRAZIL

RESOLUTION INPI/PR NO. 187, OF APRIL 27, 2017

THE PRESIDENT AND THE DIRECTOR OF PATENTS, COMPUTER PROGRAMS, AND TOPOGRAPHIES OF INTEGRATED CIRCUITS AT THE NATIONAL INSTITUTE OF INDUSTRIAL PROPERTY (INPI)
in use of the powers bestowed upon them under Decree No. 8854, of September 22, 2016,

AGREE:

Article 1
This Resolution concerns the procedures for the Submission of the Sequence Listing in electronic form, with a view to complementing the description present in the patent applications submitted before the INPI effective the date on which this Resolution entered into force, in addition to the rules for the representation of the sequences of nucleotides and/or amino acids in the Sequence Listing.

Article 2
The applicant for the patent containing one or more sequences of nucleotides and/or amino acids, which are essential to the description of the invention, shall represent them in a Sequence Listing, with a view to ensuring that the description is sufficient, pursuant to the provisions of Article 24 of Law No. 9279 of May 14, 1996 (hereinafter, LPI).

Article 3
The Sequence Listing must be presented to the INPI as an instrument that complements the description, in a computer-readable format (electronic file), burned to a non-rewritable compact disc (CD) or a non-rewritable digital disc (DVD); the electronic file containing the Sequence Listing must be saved in text format (TXT).

Article 4
The representation of the sequences of nucleotides and/or amino acids in the Sequence Listing must follow WIPO Standard ST.25, defined by the World Intellectual Property Organization (WIPO), pursuant to the rules set out in the Annex to this Resolution.

1. The Sequence Listing shall contain all linear sequences with 4 (four) or more continuous L-amino acids from a peptide or protein and all linear sequences with 10 (ten) or more continuous nucleotides, including those not included in the claim, for example, PCR probes, provided that they satisfy the conditions set out in this paragraph.
2. Branched sequences, sequences with fewer than 10 (ten) nucleotides, sequences with fewer than 4 (four) L-amino acids, and amino-acid sequences that contain at least one D-amino acid, in addition to sequences containing nucleotides or amino acids other than those listed in Tables 1, 2, 3, and 4, in the Annex to this Resolution, must be included in the patent application's description and shall not be included in the Sequence Listing.

Article 5

The CD or DVD submitted, containing the electronic Sequence Listing file in TXT format, shall also contain an electronic file corresponding to the Alphanumeric Control Code for the Sequence Listing file in TXT format, in order to certify the authenticity of its content.
1. The electronic file containing the Alphanumeric Control Code for the Sequence Listing shall be generated automatically using the Sequence Listing file in TXT format, via SisBioList, when generating the electronic Sequence Listing file in Portable Document Format (PDF).
2. Submitting the PDF file corresponding to the copy of the Sequence Listing on the CD or DVD is optional.

Article 6

The CD or DVD containing the electronic Sequence Listing file in TXT format and the electronic file containing the Alphanumeric Control Code of the Sequence Listing, must be submitted to the INPI at the time that the patent application is filed.
1. If the CD or DVD is not submitted to the INPI when the patent application is filed, it may be submitted by the applicant, regardless of any notification or request being made by the INPI, up to the date on which the patent application is examined, pursuant to Article 33 of the LPI, as part of a submission for which no payment is required.
2. Submissions made as provided for in the preceding paragraph must contain a statement from the applicant that "the information contained in the Sequence Listing submitted in electronic format is limited to the content of the material disclosed as part of the amino-acid and/or nucleotide sequences contained in the patent application, as filed".

3. If the Sequence Listing in the electronic file format is not submitted before the deadline set out in the heading and paragraph 1 of this article, the INPI will take the necessary measures to correct the patent application, with a view to satisfying the provisions of this Regulatory Instruction, which must be complied with under the terms and within the deadlines set out in the LPI.
4. In terms of satisfying the provisions of the preceding paragraph, the applicant must submit a statement declaring that "the information contained in the Sequence Listing submitted in electronic format is limited to the content of the material disclosed as part of the amino-acid and/or nucleotide sequences contained in the patent application, as filed".

Article 7
If the Sequence Listing is corrected after it is submitted, whether ex officio or at the request of the applicant, the latter shall provide the INPI with a new CD or DVD containing the corrected Sequence Listing in TXT format, pursuant to the provisions of Article 6 of this Resolution.
Sole paragraph. In terms of the provisions of the heading of this article, the applicant will be required to provide the INPI with a CD or DVD containing the corrected electronic Sequence Listing file in TXT format and the electronic file containing the Alphanumeric Control Code for the corrected Sequence Listing, as part of a submission, in addition to proof of payment of the administration fee and a statement made by the applicant that "the information contained in the corrected Sequence Listing, submitted in electronic format, does not represent an addition to the material contained in the corresponding patent application, as filed previously".

Article 8
The CD or DVD submitted containing the electronic Sequence Listing file in TXT format and the electronic file containing the Alphanumeric Control Code for the Sequence Listing, must be identified with a label, which must contain the Alphanumeric Control Code for the Sequence Listing and the Single Payment Guide (GRU) concerning the corresponding administrative action, as applicable.
Sole paragraph. If the CD or DVD submitted corresponds to a patent application that has already been filed with the INPI and that has been assigned a number, the label must also contain the patent application number.

Article 9
The CD or DVD containing the electronic Sequence Listing file in TXT format and the electronic file containing the Alphanumeric Control Code for the Sequence Listing, must be submitted in a CD or DVD case.

Article 10
In terms of the presentation of the CD or DVD containing the electronic Sequence Listing file, under the terms and within the deadlines set out in this Resolution, the patent applicant must inform the INPI, in the dedicated field of the corresponding form, that they are submitting the Sequence Listing, in addition to the Alphanumeric Control Code for the Sequence Listing, as indicated in the form itself.

Article 11
The provisions of this Resolution apply to patent applications generated as part of international patent applications filed under the terms of the Patent Cooperation Treaty (PCT), when they enter the national phase, filed before the INPI, pursuant to the legislation in force.

Article 12
The Sequence Listing may also be submitted in printed format, as an integral part of the patent application.
1. Any Sequence Listing submitted additionally in printed format when filing the patent application must be placed after the description, starting on a separate page under the title Sequence Listing.
2. The Sequence Listing pages addressed in the header must be numbered sequentially and separately, using Arabic numerals, center-aligned at the top of the page, between 1 and 2 cm from the page border.
3. The Sequence Listing referred to in the header must contain the exact same information as the content provided in the TXT format file, with the exception of the numbering of the corresponding pages, and must be accompanied by a statement by the applicant that "the Sequence Listing submitted in printed format is identical to the content of the electronic file, with the exception of page numbering".

Article 13
The Alphanumeric Control Code for the Sequence Listing shall be accompanied by a letters patent, in addition to the information and documents indicated in Article 39 of the LPI.
Sole paragraph. The Sequence Listing referred to in the header can

be accessed on the INPI's website.

Article 14

Authors of patent applications filed with the INPI before February 8, 2010, when Resolution No. 228/09 came into force requiring the submission of the Sequence Listing in electronic format, may submit the Sequence Listing in electronic format, under the conditions set out in this Resolution, voluntarily, by means of a submission for which no payment is required.

Article 15

Authors of patent applications filed with the INPI before February 8, 2010, when Resolution No. 228/09 came into force requiring the submission of the Sequence Listing in electronic format, and who submitted the Sequence Listing in breach of the legislation in force when the application was filed, must submit the Sequence Listing in electronic format, under the conditions set out in this Resolution, at the request of the INPI, unless the prerogative indicated in Article 14 is applied, attaching proof of payment of the administration fee.

Article 16

Authors of patent applications may submit the Sequence Listing in electronic format, pursuant to the specific rules set out to this end on the INPI website.

Article 17

This Resolution shall come into force on the date it is published in the Industrial Property Electronic Gazette.

Rio de Janeiro, April 27, 2017

ANNEX TO INPI/PR RESOLUTION NO. 187, OF APRIL 27, 2017

RULES FOR THE PRESENTATION AND REPRESENTATION OF AMINO-ACID AND NUCLEOTIDE SEQUENCES IN THE "SEQUENCE LISTING" IN WIPO ST.25 FORMAT

1. Definitions:

1.1 Sequence identifier is a unique integer that corresponds to the SEQ ID NO assigned to each sequence in the sequence listing, with the first sequence defined in the "Sequence Listing" corresponding to SEQ ID NO: 1, which must be the invention's most important sequence.

1.2 Numeric identifier is a three-digit number which represents a specific data element, housed between < > symbols.

1.3 Language-neutral vocabulary is a controlled vocabulary used in the sequence listing that represents scientific terms as prescribed by sequence database providers (including scientific names, qualifiers, and their controlled-vocabulary values, the symbols appearing in Tables 1, 2, 3, and 4, and the feature keys appearing in Tables 5 and 6);

1.4 "Free text" is a wording describing characteristics of the sequence under the numeric identifier (Other information), which does not use the language-neutral vocabulary as referred to in paragraph 1.3.

2. Representation of biological sequences in WIPO ST.25 format:

2.1 Each sequence shall be assigned a separate sequence identifier. The sequence identifiers shall begin with 1 and increase sequentially by integers, for example: "SEQ ID NO:1", "SEQ ID NO:2", "SEQ ID NO:3", etc.

2.2 In the description, in the claims or drawings of the application, the sequences represented in the sequence listing shall be referred to by the sequence identifier and preceded by "SEQ ID NO:".

2.3 Nucleotide and amino acid sequences should be represented by at least one of the following three possibilities: (i) a pure nucleotide sequence; (ii) a pure amino acid sequence; (iii) a nucleotide sequence together with its corresponding amino acid sequence.

2.4 For those sequences represented in the format specified in option (iii), the amino acid sequence must be disclosed separately in the sequence listing as a pure amino acid sequence with a separate integer sequence identifier.

3. The format and symbols to be used for nucleotide sequences:

3.1 A nucleotide sequence shall be represented only by a single strand, in the 5'-end to 3'-end direction from left to right.

3.2 A nucleotide sequence shall be listed with a maximum of 60 bases per line, with a space between each group of 10 bases.

3.3 The bases of the coding regions of a nucleotide sequence shall be listed as triplets (codons).

3.4 The bases of a nucleotide sequence shall be listed using a one-letter code for nucleotide sequence characters. Only lower case letters in conformity with the list given in Table 1 shall be used.

3.5 Modified bases shall be represented as the corresponding unmodified bases or as "n" in the sequence itself if the modified base is one of those listed in Table 2.

4. The format and symbols to be used for amino acid sequences:

4.1 A protein or peptide sequence shall be listed with a maximum of 16 amino acids per line, with a space provided between each amino acid.

4.2 Amino acids corresponding to the codons in the coding parts of a nucleotide sequence shall be placed immediately under the corresponding codons. Where a codon is split by an intron, the amino acid symbol should be given below the portion of the codon containing two nucleotides.

4.3 The enumeration of amino acids shall start at the first amino acid of the sequence, with number 1.

4.4 Optionally, the amino acids receding the mature protein, for example pre-sequences, pro-sequences, pre-pro-sequences, and signal sequences, when present, may have negative numbers, counting backwards starting with the amino acid next to number 1.

4.5 Zero (0) is not used when the numbering of amino acids uses negative numbers to distinguish the mature protein.

4.6 An amino acid sequence that is made up of one or more non-contiguous segments of a larger sequence or of segments from different sequences shall be numbered as a separate sequence, with a separate sequence identifier.

4.7 The amino acids in a protein or peptide sequence shall be listed in the amino to carboxy direction from left to right. The amino and carboxy groups shall not be represented in the sequence.

4.8 The amino acids shall be represented using the three-letter code with the first letter as a capital and shall conform to the list given in Table 3.

5. Mandatory data elements:

5.1 The sequence listing shall include, in addition to and immediately preceding the actual nucleotide and/or amino acid sequence, the following items of information (mandatory data elements):

<110>	**Applicant name**
<120>	**Title of invention**
<160>	**Total number of SEQ ID NOs**
<210>	**SEQ ID NO: #**
<211>	**Length**
<212>	**Type**
<213>	**Organism**
<400>	**Sequence**

Where the name of the applicant (numeric identifier) is written in characters other than those of the Latin alphabet, it shall also be indicated in characters of the Latin alphabet either as a mere transliteration or through translation into English.

5.2 If "n" or "Xaa" or a modified base or modified/unusual L-amino acid is used in the sequence, the following data elements are mandatory:

<220>	**Feature**
<221>	**Name/key**
<222>	**Location**
<223>	**Other information**

5.3 If the organism (numeric identifier) is "Artificial Sequence" or "Unknown," the following data elements are mandatory:

<220>	**Feature**
<223>	**Other information**

5.4 When a sequence listing is submitted when the patent application is filed or at any time prior to the assignment of an application number, the following data element must be included in the sequence listing:

<130>	**Reference number (designated by the applicant)**

5.5 When a sequence listing is submitted in response to a request made by the INPI or at any time following the assignment of an application number, the following data elements must be included in the "Sequence Listing":

<140>	**Current patent application number**
<141>	**Current filing date**

5.6 In addition to the data elements identified above, when a sequence listing is filed relating to an application which claims the priority of an earlier application, the following data elements shall be included in the "Sequence Listing":

<150>	**Earlier patent application (priority document)**
<151>	**Earlier filing date (day/month/year)**

6. Presentation of features:

6.1 When features of sequences are presented (that is, numeric identifier <220>), they shall be described by the "feature keys" set out in Tables 5 and 6.

7. Free text:

7.1 The use of free text shall be limited to a few short terms that are essential for understanding the sequence.

7.2 Each data element shall not exceed four lines with a maximum of 65 characters per line.

7.3 Any further information shall be included in the main part of the description.

Mandatory Numeric Identifiers

Numeric Identifier	Numeric Identifier Description	Comment
<110>	Applicant name	Where the name of the applicant is written in characters other than those of the Latin alphabet, the same shall also be indicated in characters of the Latin alphabet either as a mere transliteration or through translation into English; if there is more than one applicant, introduce one name per line
<120>	Title of invention	In native language
<130>	Application reference number	Only mandatory under the conditions specified in paragraph 5.4.
<140>	Current patent application	Only mandatory under the conditions specified in paragraph 5.5.
<141>	Current filing date	Only mandatory under the conditions specified in paragraph 5.5.
<150>	Earlier patent application (priority)	Only mandatory under the conditions specified in paragraph 5.6.
<151>	Earlier filing date (priority)	Only mandatory under the conditions specified in paragraph 5.6.
<160>	Number of SEQ ID NOs	Includes the number of SEQ ID NOs contained in the sequence listing
<210>	Information on the SEQ ID NO: #	The response shall be an integer representing the SEQ ID NO shown
<211>	Length	Sequence length expressed in number of base pairs or amino acids
<212>	Type	Type of DNA/RNA/PROTEIN molecule sequenced in SEQ ID NO: #, either DNA, RNA, or PRT (protein); if a nucleotide sequence contains both DNA and RNA fragments, the value shall be "DNA"; in addition, the combined DNA/RNA molecule shall be further described in

		the section of features **<220>** to **<223>**
<213>	Organism	Genus and species (that is, scientific name) or "Artificial Sequence" or "Unknown"; in addition, the artificial sequence or unknown organism must also be described in the section of features **<220>** to **<223>**
<220>	Feature	Only mandatory for the conditions specified in paragraphs 5.2 and 5.3. Otherwise, leave empty.
<221>	Name/key	Only mandatory for the condition specified in paragraph 5.2.
<222>	Location	Only mandatory for the condition specified in paragraph 5.2.
<223>	Other information	Only mandatory for the conditions specified in paragraphs 5.2 and 5.3.
<400>	Sequence	SEQ ID NO: should follow the numeric identifier and should appear on the line preceding the sequence in question

Table 1: List of 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

Table 2: 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

Table 3: List of Amino Acids

Symbol	**Meaning**
Ala	Alanine
Cys	Cysteine
Asp	Aspartic Acid
Glu	Glutamic Acid
Phe	Phenylalanine
Gly	Gylcine
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

Table 4: List of Modified or 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
EtGl	N-Ethylglycine
EtAsn	N-Ethylasparagine
Hyl	Hydroxylysine
aHyl	allo-Hydroxylysine
3Hyp	3-Hydroxyproline
4Hyp	4-Hydroxyproline
Ide	Isodesmosine
alle	allo-Isoleucine
MeGly	N-Methylglycine, sarcosine
Melle	N-Methylisoleucine
MeLys	6-N-Methyllysine
MeVal	N-Methylvaline
Nva	Norvaline
Nle	Norleucine
Orn	Ornithine

Table 5: List of Feature Keys Related to Nucleotide Sequences

Key	Description
allele	a related individual or strain containing 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 in the same DNA strand and can function in either orientation and in any location (5' or 3') relative to the promoter
exon	region of genome that codes for portion of spliced mRNA; may contain 5'UTR, all 15 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 (5' or 3'); 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 (5' and 3') 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, 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 poly adenylation
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 permissible
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_segment, D_segment, N_region and J_segment
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

Table 6: List of Feature Keys Related to Amino Acid 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 alphacarboxyl 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
TRANSMEM	extent of a DNA-binding 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	nonconsecutive 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

8. Optional data elements:

8.1 All data elements mentioned below are optional in the "Sequence Listing":

<170>	Software used to generate the sequence listing
<300>	Publication information; if there are several publications, repeat the section for each relevant publication
<301>	Authors, provide one name per line, preferentially in the following format: surname, other names, and/or initials
<302>	Title of publication
<303>	journal name in which data published
<304>	journal volume in which data published
<305>	journal issue number in which data published
<306>	journal page numbers on which data published
<307>	journal date on which data published; Use Day/Month/Year format
<308>	accession number assigned by database including database name
<309>	date of entry in database (day/month/year)
<310>	document number, for patent type citations only

<311>	document filing date, for patent type citations only (day/month/year)
<312>	document publication date; for patent type citations only (day/month/year)
<313>	Relevant residues in SEQ ID NO: #: FROM_TO