STANDARD ST.25

STANDARD FOR THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS IN PATENT APPLICATIONS

Standard adopted by the PCIPI Executive Coordination Committee at its twenty-second session on May 28, 1998

It is recommended that Offices apply the provisions of the "Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings in International Patent Applications Under the Patent Cooperation Treaty (PCT)" as set out in Annex C to the Administrative Instructions under the PCT, *mutatis mutandis*, to all patent applications other than the PCT international applications, noting that certain provisions specific to PCT procedures and requirements may not be applicable to patent applications other than PCT international applications(*). The text of that PCT Standard is reproduced on the following pages.

(*) If, on July 1, 1998, the national law and practice applicable by an Office is not compatible with the provisions of the first two sentences of paragraph 3 of the "Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings in International Patent Applications Under the Patent Cooperation Treaty (PCT)," that Office may choose not to follow those provisions for as long as that incompatibility continues.

ANNEX C

STANDARD FOR THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS IN INTERNATIONAL PATENT APPLICATIONS UNDER THE PCT

INTRODUCTION

1. This Standard has been elaborated so as to provide standardization of the presentation of nucleotide and amino acid sequence listings in international patent applications. The Standard is intended to allow the applicant to draw up a single sequence listing which is acceptable to all receiving Offices, International Searching and Preliminary Examining Authorities for the purposes of the international phase, and to all designated and elected Offices for the purposes of the national phase. It is intended to enhance the accuracy and quality of presentations of nucleotide and amino acid sequences given in international applications, to make for easier presentation and dissemination of sequences for the benefit of applicants, the public and examiners, to facilitate searching of sequence data and to allow the exchange of sequence data in electronic form and the introduction of sequence data onto computerized databases.

DEFINITIONS

2. For the purposes of this Standard:

(i) the expression “sequence listing” means a part of the description of the application as filed or a document filed subsequently to the application, which gives a detailed disclosure of the nucleotide and/or amino acid sequences and other available information;

(ii) sequences which are included are any unbranched sequences of four or more amino acids or unbranched sequences of ten or more nucleotides. Branched sequences, sequences with fewer than four specifically defined nucleotides or amino acids as well as sequences comprising nucleotides or amino acids other than those listed in Appendix 2, Tables 1, 2, 3 and 4, are specifically excluded from this definition;

(iii) “nucleotides” embrace only those nucleotides that can be represented using the symbols set forth in Appendix 2, Table 1. Modifications, for example, methylated bases, may be described as set forth in Appendix 2, Table 2, but shall not be shown explicitly in the nucleotide sequence;

(iv) “amino acids” are those L-amino acids commonly found in naturally occurring proteins and are listed in Appendix 2, Table 3. Those amino acid sequences containing at least one D-amino acid are not intended to be embraced by this definition. Any amino acid sequence that contains post-translationally modified amino acids may be described as the amino acid sequence that is initially translated using the symbols shown in Appendix 2, Table 3, with the modified positions, for example, hydroxylations or glycosylations, being described as set forth in Appendix 2, Table 4, but these modifications shall not be shown explicitly in the amino acid sequence. Any peptide or protein that can be expressed as a sequence using the symbols in Appendix 2, Table 3, in conjunction with a description elsewhere to describe, for example, abnormal linkages, cross-links (for example, disulfide bridge) and end caps, non-peptidyl bonds, etc., is embraced by this definition;

(v) “sequence identifier” is a unique integer that corresponds to the SEQ ID NO assigned to each sequence in the listing;

(vi) “numeric identifier” is a three-digit number which represents a specific data element;

(vii) “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 Appendix 2, Tables 1, 2, 3 and 4, and the feature keys appearing in Appendix 2, Tables 5 and 6;

(viii) “competent Authority” is the International Searching Authority that is to carry out the international search on the international application, or the International Preliminary Examining Authority that is to carry out the international preliminary examination on the international application, or the designated/elected Office before which the processing of the international application has started.

SEQUENCE LISTING

3. The sequence listing as defined in paragraph 2(i) shall, where it is filed together with the application, be placed at the end of the application. This part shall be entitled “Sequence Listing,” begin on a new page and preferably have independent page numbering. The sequence listing forms an integral part of the description; it is therefore unnecessary, subject to paragraph 36, to describe the sequences elsewhere in the description.

4. Where the sequence listing as defined in paragraph 2(i) is not contained in the application as filed but is a separate document furnished subsequently to the filing of the application (see paragraph 37), it shall be entitled “Sequence Listing” and shall have independent page numbering. The original numbering of the sequences (see paragraph 5) in the application as filed shall be maintained in the subsequently furnished sequence listing.

5. Each sequence shall be assigned a separate sequence identifier. The sequence identifiers shall begin with 1 and increase sequentially by integers. If no sequence is present for a sequence identifier, the code 000 should appear under numeric identifier <400>, beginning on the next line following the SEQ ID NO. The response for numeric identifier <160> shall include the total number of SEQ ID NOs, whether followed by a sequence or by the code 000.

6. In the description, 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:”.

7. 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.

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

NUCLEOTIDE SEQUENCES

Symbols to Be Used

8. A nucleotide sequence shall be presented only by a single strand, in the 5’-end to 3’-end direction from left to right. The terms 3’ and 5’ shall not be represented in the sequence.

9. The bases of a nucleotide sequence shall be represented using the one-letter code for nucleotide sequence characters. Only lower case letters in conformity with the list given in Appendix 2, Table 1, shall be used.

10. 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 Appendix 2, Table 2, and the modification shall be further described in the feature section of the sequence listing, using the codes given in Appendix 2, Table 2. These codes may be used in the description or the feature section of the sequence listing but not in the sequence itself (see also paragraph 32). The symbol “n” is the equivalent of only one unknown or modified nucleotide.

Format to Be Used

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

12. The bases of a nucleotide sequence (including introns) shall be listed in groups of 10 bases, except in the coding parts of the sequence. Leftover bases, fewer than 10 in number at the end of non-coding parts of a sequence, should be grouped together and separated from adjacent groups by a space.

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

14. The enumeration of the nucleotide shall start at the first base of the sequence with number 1. It shall be continuous through the whole sequence in the direction 5' to 3'. It shall be marked in the right margin, next to the line containing the one-letter codes for the bases, and giving the number of the last base of that line. The enumeration method for nucleotide sequences set forth above remains applicable to nucleotide sequences that are circular in configuration, with the exception that the designation of the first nucleotide of the sequence may be made at the option of the applicant.

15. A nucleotide 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. A sequence with a gap or gaps shall be numbered as a plurality of separate sequences with separate sequence identifiers, with the number of separate sequences being equal in number to the number of continuous strings of sequence data.

AMINO ACID SEQUENCES

Symbols to Be Used

16. 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.

17. 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 Appendix 2, Table 3. An amino acid sequence that contains a blank or internal terminator symbols (for example, "Ter" or "*" or ".") may not be represented as a single amino acid sequence, but shall be presented as separate amino acid sequences (see paragraph 22).

18. Modified and unusual amino acids shall be represented as the corresponding unmodified amino acids or as "Xaa" in the sequence itself if the modified amino acid is one of those listed in Appendix 2, Table 4, and the modification shall be further described in the feature section of the sequence listing, using the codes given in Appendix 2, Table 4. These codes may be used in the description or the feature section of the sequence listing but not in the sequence itself (see also paragraph 32). The symbol "Xaa" is the equivalent of only one unknown or modified amino acid.

Format to Be Used

19. 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.

20. 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.

21. The enumeration of amino acids shall start at the first amino acid of the sequence, with number 1. Optionally, the amino acids preceding 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. Zero (0) is not used when the numbering of amino acids uses negative numbers to distinguish the mature protein. It shall be marked under the sequence every five amino acids. The enumeration method for amino acid sequences set forth above remains applicable for amino acid sequences that are circular in configuration, with the exception that the designation of the first amino acid of the sequence may be made at the option of the applicant.

22. 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. A sequence with a gap or gaps shall be numbered as a plurality of separate sequences with separate sequence identifiers, with the number of separate sequences being equal in number to the number of continuous strings of sequence data.

OTHER AVAILABLE INFORMATION IN THE SEQUENCE LISTING

23. The order of the items of information in the sequence listings shall follow the order in which those items are listed in the list of numeric identifiers of data elements as defined in Appendix 1.

24. Only numeric identifiers of data elements as defined in Appendix 1 shall be used for the presentation of the items of information in the sequence listing. The corresponding numeric identifier descriptions shall not be used. The provided information shall follow immediately after the numeric identifier while only those numeric identifiers for which information is given need appear on the sequence listing. Two exceptions to this requirement are numeric identifiers <220> and <300>, which serve as headers for "Feature" and "Publication Information," respectively, and are associated with information in numeric identifiers <221> to <223> and <301> to <313>, respectively. When feature and publication information is provided in the sequence listing under those numeric identifiers, numeric identifiers <220> and <300>, respectively, should be included, but left blank. Generally, a blank line shall be inserted between numeric identifiers when the digit in the first or second position of the numeric identifier changes. An exception to this general rule is that no blank line should appear preceding numeric identifier <310>. Additionally, a blank line shall precede any repeated numeric identifier.

Mandatory Data Elements

25. The sequence listing shall include, in addition to and immediately preceding the actual nucleotide and/or amino acid sequence, the following items of information defined in Appendix 1 (mandatory data elements):

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

Where the name of the applicant (numeric identifier <110>) 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.

The data elements, except those under numeric identifiers <110>, <120> and <160>, shall be repeated for each sequence included in the sequence listing. Only the data elements under numeric identifiers <210> and <400> are mandatory if no sequence is present for a sequence identifier (see paragraph 5, above, and SEQ ID NO: 4 in the example depicted in Appendix 3 of this Standard).

26. In addition to the data elements identified in paragraph 25, above, when a sequence listing is filed at the same time as the application to which it pertains or at any time prior to the assignment of an application number, the following data element shall be included in the sequence listing:

<130>	File reference

27. In addition to the data elements identified in paragraph 25, above, when a sequence listing is filed in response to a request from a competent Authority or at any time following the assignment of an application number, the following data elements shall be included in the sequence listing:

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

28. In addition to the data elements identified in paragraph 25, 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
<151>	Earlier application filing date

29. 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

30. If the organism (numeric identifier <213>) is “Artificial Sequence” or “Unknown,” the following data elements are mandatory:

<220>	Feature
<223>	Other information

Optional Data Elements

31. All data elements defined in Appendix 1, not mentioned in paragraphs 25 to 30, above, are optional (optional data elements).

Presentation of Features

32. When features of sequences are presented (that is, numeric identifier <220>), they shall be described by the “feature keys” set out in Appendix 2, Tables 5 and 6.[(1)]

Free Text

33. “Free text” is a wording describing characteristics of the sequence under numeric identifier <223> (Other information) which does not use language-neutral vocabulary as referred to in paragraph 2(vii).

34. The use of free text shall be limited to a few short terms indispensable for the understanding of the sequence. It shall not exceed four lines with a maximum of 65 characters per line for each given data element, when written in English. Any further information shall be included in the main part of the description in the language thereof.

35. Any free text should preferably be in the English language.

36. Where the sequence listing part of the description contains free text, any such free text shall be repeated in the main part of the description in the language thereof. It is recommended that the free text in the language of the main part of the description be put in a specific section of the description called “Sequence Listing Free Text.”

SUBSEQUENTLY FURNISHED SEQUENCE LISTING

37. Any sequence listing which is not contained in the application as filed but which is furnished subsequently shall not go beyond the disclosure in the application as filed and shall be accompanied by a statement to that effect. This means that a sequence listing furnished subsequently to the filing of the application shall contain only those sequences that were disclosed in the application as filed.

38. Any sequence listing not contained in the application as filed does not form part of the application. However, the provisions of PCT Rules 13*ter*, 26.3 and 91 and PCT Article 34 would apply, so that it may be possible, subject to the applicable provisions, for a sequence listing contained in the application as filed to be corrected under PCT Rules 13*ter* or 26.3, rectified under PCT Rule 91 (in the case of an obvious error), or amended under PCT Article 34, or for a sequence listing to be submitted under PCT Article 34 as an amendment to the application.

(1) These tables contain extracts of the DDBJ/EMBL/GenBank Feature Table (nucleotide sequences) and the SWISS PROT Feature Table (amino acid sequences).

COMPUTER READABLE FORM OF THE SEQUENCE LISTING

39. A copy of the sequence listing shall also be submitted in computer readable form, in addition to the sequence listing as contained in the application, whenever this is required by the competent Authority.

40. Any sequence listing in computer readable form submitted in addition to the written sequence listing shall be identical to the written sequence listing and shall be accompanied by a statement that "the information recorded in computer readable form is identical to the written sequence listing."

41. The entire printable copy of the sequence listing shall be contained within one electronic file preferably on a single diskette or any other electronic medium that is acceptable to the competent Authority. The file recorded on the diskette or any other electronic medium that is acceptable to the competent Authority shall be encoded using IBM[(2)] Code Page 437, IBM Code Page 932[(3)] or a compatible code page. A compatible code page, as would be required for, for example, Japanese, Chinese, Cyrillic, Arabic, Greek or Hebrew characters, is one that assigns the Roman alphabet and numerals to the same hexadecimal positions as do the specified code pages.

42. The computer readable form shall preferably be created by dedicated software such as PatentIn or other custom computer programs; it may be created by any means, as long as the sequence listing on a submitted diskette or any other electronic medium that is acceptable to the competent Authority is readable under a Personal Computer Operating system that is acceptable to the competent Authority.

43. File compression is acceptable when using diskette media, so long as the compressed file is in a self-extracting format that will decompress on a Personal Computer Operating system that is acceptable to the competent Authority.

44. The diskette or any other electronic medium that is acceptable to the competent Authority shall have a label permanently affixed thereto on which has been hand-printed, in block capitals or typed, the name of the applicant, the title of the invention, a reference number, the date on which the data were recorded, the computer operating system and the name of the competent Authority.

45. If the diskette or any other electronic medium that is acceptable to the competent Authority is submitted after the date of filing of an application, the labels shall also include the filing date of the application and the application number.

46. Any correction of the written sequence listing which is submitted under PCT Rules 13*ter*.1(a)(i) or 26.3, any rectification of an obvious error in the written sequence listing which is submitted under PCT Rule 91, or any amendment which includes a written sequence listing and which is submitted under PCT Article 34, shall be accompanied by a computer readable form of the sequence listing including any such correction, rectification or amendment.

[Appendices 1 to 3 follow]

(2) IBM is a registered trademark of International Business Machine Corporation, United States of America.
(3) The specified code pages are *de facto* standards for personal computers.

APPENDICES

APPENDIX 1

NUMERIC IDENTIFIERS

Only numeric identifiers as defined below may be used in sequence listings submitted in applications. The text of the data element headings given below shall not be included in the sequence listings.

Numeric identifiers of mandatory data elements, that is, data elements which must be included in all sequence listings (see paragraph 25 of this Standard: items 110, 120, 160, 210, 211, 212, 213 and 400) and numeric identifiers of data elements which must be included in circumstances specified in this Standard (see paragraphs 26, 27, 28, 29 and 30 of this Standard: items 130, 140, 141, 150 and 151, and 220 to 223) are marked by the symbol “M.”

Numeric identifiers of optional data elements (see paragraph 31 of this Standard) are marked by the symbol ”O.”

Numeric Identifier	Numeric Identifier Description	Mandatory (M) or Optional (O)	Comment
<110>	Applicant name	M	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
<120>	Title of invention	M	
<130>	File reference	M, in the circumstances specified in paragraph 26 of this Standard	see paragraph 26 of this Standard
<140>	Current patent application	M, in the circumstances specified in paragraph 27 of this Standard	see paragraph 27 of this Standard; the current patent application shall be identified, in the following order, by the two-letter code indicated in accordance with WIPO Standard ST.3 and the application number (in the format used by the industrial property Office with which the current patent application is filed) or, for an international application, by the international application number
<141>	Current filing date	M, in the circumstances specified in paragraph 27 of this Standard	see paragraph 27 of this Standard; the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)

Numeric Identifier	Numeric Identifier Description	Mandatory (M) or Optional (O)	Comment
<150>	Earlier patent application	M, in the circumstances specified in paragraph 28 of this Standard	see paragraph 28 of this Standard; the earlier patent application shall be identified, in the following order, by the two-letter code indicated in accordance with WIPO Standard ST.3 and the application number (in the format used by the industrial property Office with which the earlier patent application was filed) or, for an international application, by the international application number
<151>	Earlier application filing date	M, in the circumstances specified in paragraph 28 of this Standard	see paragraph 28 of this Standard; the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)
<160>	Number of SEQ ID NOs	M	
<170>	Software	O	
<210>	Information for SEQ ID NO: x	M	response shall be an integer representing the SEQ ID NO shown
<211>	Length	M	sequence length expressed in number of base pairs or amino acids
<212>	Type	M	type of molecule sequenced in SEQ ID NO: x, either DNA, RNA or PRT; 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 <220> to <223> feature section
<213>	Organism	M	Genus Species (that is, scientific name) or “Artificial Sequence” or “Unknown”
<220>	Feature	M, in the circumstances specified in paragraph 29 and 30 of this Standard	leave blank; see paragraphs 29 and 30 of this Standard; description of points of biological significance in the sequence in SEQ ID NO: x) (may be repeated depending on the number of features indicated)
<221>	Name/key	M, in the circumstances specified in paragraph 29 of this Standard	see paragraph 29 of this Standard; only those keys as described in Table 5 or 6 of Appendix 2 shall be used

Numeric Identifier	Numeric Identifier Description	Mandatory (M) or Optional (O)	Comment
<222>	Location	M, in the circumstances specified in paragraph 29 of this Standard	see paragraph 29 of this Standard; – from (number of first base/amino acid in the feature) – to (number of last base/amino acid in the feature) – base pairs (numbers refer to positions of base pairs in a nucleotide sequence) – amino acids (numbers refer to positions of amino acid residues in an amino acid sequence) – whether feature is located on the complementary strand to that filed in the sequence listing
<223>	Other information:	M, in the circumstances specified in paragraphs 29 and 30 of this Standard	see paragraphs 29 and 30 of this Standard; any other relevant information, using language neutral vocabulary, or free text (preferably in English); any free text is to be repeated in the main part of the description in the language thereof (see paragraph 36 of this Standard); where any modified base or modified/unusual L-amino acid appearing in Appendix 2, Tables 2 and 4, is in the sequence, the symbol associated with that base or amino acid from Appendix 2, Tables 2 and 4, should be used
<300>	Publication information	O	leave blank; repeat section for each relevant publication
<301>	Authors	O	
<302>	Title	O	title of publication
<303>	Journal	O	journal name in which data published
<304>	Volume	O	journal volume in which data published
<305>	Issue	O	journal issue number in which data published
<306>	Pages	O	journal page numbers on which data published
<307>	Date	O	journal date on which data published; if possible, the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)
<308>	Database accession number	O	accession number assigned by database including database name
<309>	Database entry date	O	date of entry in database; the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)

Numeric Identifier	Numeric Identifier Description	Mandatory (M) or Optional (O)	Comment
<310>	Document number	O	document number, for patent type citations only; the full document shall specify, in the following order, the two-letter code indicated in accordance with WIPO Standard ST.3, the publication number indicated in accordance with WIPO Standard ST.6, and the kind-of-document code indicated in accordance with WIPO Standard ST.16
<311>	Filing date	O	document filing date, for patent-type citations only; the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)
<312>	Publication date	O	document publication date; for patent-type citations only; the date shall be indicated in accordance with WIPO Standard ST.2 (CCYY MM DD)
<313>	Relevant residues in SEQ ID NO: x: from to	O	
<400>	Sequence	M	SEQ ID NO: x should follow the numeric identifier and should appear on the line preceding the sequence (see Appendix 3)

[Appendix 2 follows]

APPENDIX 2

NUCLEOTIDE AND AMINO ACID SYMBOLS AND FEATURE TABLE

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	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

Table 4: 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

Table 5: 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

Key	Description
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

Key	Description
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

Key	Description
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 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

Key	Description
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

[Appendix 3 follows]

APPENDIX 3

SPECIMEN SEQUENCE LISTING

```
<110>       Smith, John;  Smithgene Inc.

<120>       Example of a Sequence Listing

<130>       01-00001

<140>       PCT/EP98/00001
<141>       1998-12-31

<150>       US 08/999,999
<151>       1997-10-15

<160>       4

<170>       PatentIn version 2.0

<210>       1
<211>       389
<212>       DNA
<213>       Paramecium sp.

<220>
<221>       CDS
<222>       (279)...(389)

<300>
<301>       Doe, Richard
<302>       Isolation and Characterization of a Gene Encoding a Protease
            from Paramecium sp.
<303>       Journal of Genes
<304>       1
<305>       4
<306>       1-7
<307>       1988-06-31
<308>       123456
<309>       1988-06-31

<400>       1
agctgtagtc  attcctgtgt  cctcttctct  ctgggcttct  caccctgcta  atcagatctc      60

agggagagtg  tcttgaccct  cctctgcctt  tgcagcttca  caggcaggca  ggcaggcagc      120

tgatgtggca  attgctggca  gtgccacagg  cttttcagcc  aggcttaggg  tgggttccgc      180

cgcggcgcgg  cggcccctct  cgcgctcctc  tcgcgcctct  ctctcgctct  cctctcgctc      240
```

```
ggacctgatt   aggtgagcag   gaggaggggg   cagttagc     atg  gtt  tca  atg  ttc  agc   296
                                                    Met  Val  Ser  Met  Phe  Ser
                                                     1                   5

ttg  tct  ttc  aaa  tgg  cct  gga  ttt  tgt  ttg  ttt  gtt  tgt  ttg  ttc  caa   344
Leu  Ser  Phe  Lys  Trp  Pro  Gly  Phe  Cys  Leu  Phe  Val  Cys  Leu  Phe  Gln
               10                        15                  20

tgt  ccc  aaa  gtc  ctc  ccc  tgt  cac  tca  tca  ctg  cag  ccg  aat  ctt       389
Cys  Pro  Lys  Val  Leu  Pro  Cys  His  Ser  Ser  Leu  Gln  Pro  Asn  Leu
          25                        30                  35
```

<210> 2
<211> 37
<212> PRT
<213> Paramecium sp.

<400> 2

```
Met  Val  Ser  Met  Phe  Ser  Leu  Ser  Phe  Lys  Trp  Pro  Gly  Phe  Cys  Leu
 1                   5                        10                  15

Phe  Val  Cys  Leu  Phe  Gln  Cys  Pro  Lys  Val  Leu  Pro  Cys  His  Ser  Ser
               20                        25                  30

Leu  Gln  Pro  Asn  Leu
          35
```

<210> 3
<211> 11
<212> PRT
<213> Artificial Sequence

<220>
<223> Designed peptide based on size and polarity to act as a linker between the alpha and beta chains of Protein XYZ.

<400> 3

```
Met  Val  Asn  Leu  Glu  Pro  Met  His  Thr  Glu  Ile
 1                   5                        10
```

<210> 4
<400> 4
000

[End of Appendix 3 and of Standard]