NOTICE: This file contains information that was previously published in the WIPO *Handbook on Industrial Property Information and Documentation*, but that has become outdated.

STANDARD ST.23

RECOMMENDATION FOR THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS IN PATENT APPLICATIONS AND IN PUBLISHED PATENT DOCUMENTS

INTRODUCTION

1. This Recommendation has been elaborated so as to provide standardization of the presentation of nucleotide and amino acid sequence listings in patent applications and in published patent documents. The Recommendation is intended to reduce the cost of converting applications into a standardized digitized form for publication, to enhance the accuracy and quality of presentations of nucleotide and amino acid sequence listings given in patent applications and in patent documents, to make for easier interpretation of sequences by applicants, the public and examiners and to improve the use of computerized databases in the field of biotechnology.

2. This Recommendation has been based upon corresponding regulations already introduced by some industrial property offices. The Recommendation makes no specific provision for the presentation by applicants of nucleotide and amino acid sequence listings on electronic data carriers. Such provisions are contained in WIPO Standard ST.24 (Recommendation Concerning the Filing of Nucleotide and Amino Acid Sequence Listings in Computer-Readable Form). Failing the filing of disclosures of those sequence listings on electronic data carriers, applicants are recommended to submit typed documents that closely follow WIPO Standard ST.22 (Recommendation for the Presentation of Patent Applications Typed in Optical Character Recognition (OCR) Format) so as to permit the accurate transcription of nucleotide and amino acid sequence listings.

DEFINITIONS

3. For the purposes of this Recommendation:

(i) the expression “patent application” means a document filed by an applicant, or by an agent (representative) on his(her) behalf, requesting the grant of a patent. It usually contains a detailed description of the invention, the claims, and drawings when necessary for the understanding of the invention;

(ii) the expression “patent document” means patents for invention, utility certificates, utility models, patents or certificates of addition, utility certificates of addition, and published applications therefor;

(iii) the expression “Sequence Listing” means a last, separate part of the description in the typed patent application, which gives a detailed disclosure of the nucleotide and/or amino acid sequences and other available information. Each sequence disclosed appears separately in the Sequence Listing and is assigned a separate identifier number.

(iv) the expressions “nucleotide sequence” and “amino acid sequence” mean an unbranched sequence of ten or more contiguous nucleotides and an unbranched sequence of four or more contiguous amino acids, respectively. Branched sequences are specifically excluded from this definition. Nucleotides and amino acids are further defined as follows:

(a) “Nucleotides” embrace only those nucleotides that can be represented using the symbols set forth in paragraph 8. Modifications, e.g. methylated bases, may be described as set forth in paragraph 9, but shall not be shown explicitly in the nucleotide sequence.

(b) “Amino acids” are those L-amino acids commonly found in naturally occurring proteins and are listed in paragraph 11. Those amino acid sequences containing D-amino acids 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 paragraph 11 with the modified positions, e.g.`hydroxylations or glycosylations, being described as set forth in paragraph`12, 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 paragraph 11 in conjunction with a description elsewhere to describe, for example, abnormal linkages, cross links and end caps, non-peptidyl bonds, etc., is embraced by this definition.

SEQUENCE LISTING

4. The sequences should be given as a last, separate part of the description in the patent application or published patent document, entitled “Sequence Listing,” and should begin on a new page. Each sequence disclosed should appear separately in the “Sequence Listing.” Each sequence should be assigned a separate identifier number, written as SEQ ID NO:1, SEQ`ID`NO:2, SEQ ID NO:3, etc. It is recommended that the number of sequences presented in the Sequence Listing be indicated.

5. In the description or claims of the application, the sequences represented in the Sequence Listing should be referred to by the identifier number, even if the sequence or other additional or modified representations of the sequence are embedded in the text or in the drawings accompanying the description.

6. Nucleotide and amino acid sequences should be represented by at least one of the following three possibilities:

(i) a pure nucleotide sequence;

(ii) an amino acid sequence;

(iii) a nucleotide sequence together with its corresponding amino acid sequence.

SYMBOLS TO BE USED

7. A nucleotide sequence should be presented only by a single strand, in the 5’-end to 3’-end direction from left to right.

8. The bases of a nucleotide sequence should be represented using the one-letter code for nucleotide sequence characters. It is recommended to use upper case letters in conformity with the following list:

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) or (unknown or other)	aNy

WIPO FOR OFFICIAL USE ONLY

9. Modified bases may be represented as the corresponding unmodified bases in the sequence itself if the modified base is one of those listed below and the modification is further described elsewhere in the Sequence Listing. The codes from the list below may be used in the description or the Sequence Listing but not in the sequence itself.

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 (v)
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

Otherwise, modified bases should be listed in the sequence as “N,” with further information given elsewhere in the Sequence Listing.

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

WIPO FOR OFFICIAL USE ONLY

11. The amino acids should be represented using the three-letter code with the first letter as a capital and should conform to the following:

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

12. Modified and unusual amino acids may be represented as the corresponding unmodified amino acids in the sequence itself if the modified amino acid is one of those listed below and the modification is also further described elsewhere in the Sequence Listing. The codes from the list below may be used in the description or the Sequence Listing but not in the sequence itself.

Symbol	Meaning
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bA1a	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

Otherwise, modified or unusual amino acids should be listed in the sequence as “Xaa,” with further information given elsewhere in the Sequence Listing.

FORMAT TO BE USED

13. A nucleotide sequence should be listed with a maximum of 16 codons or 60 bases per line, with a space between each codon or group of 10 bases.

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

15. The bases of a nucleotide sequence (including introns) should 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 noncoding parts of a sequence, shall be grouped together and separated from adjacent groups by a space (see Annex 2, Example 10, nucleotides 356-360).

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

17. Amino acids corresponding to the codons in the coding parts of a nucleotide sequence should be typed immediately under the corresponding codons. Where a codon is split by an intron, the amino acid symbol should be typed below the portion of the codon containing two nucleotides (see Annex 2, Example 11, position 120).

18. The enumeration of the nucleotide bases should start at the first base of the sequence with number 1. It should be continuous through the whole sequence in the direction 5’ to 3’. It should 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 base of the nucleotide sequence may be made at the option of the applicant.

19. The enumeration of amino acids should start at the first amino acid of the mature protein, with number 1. The amino acids preceding the mature protein, for example pre-sequences, pro-sequences, pre-pro-sequences and signal sequences, when present, should 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. Otherwise, the enumeration of amino acids should start at the first amino acid at the amino terminal as number 1. It should be marked under the sequence every 5 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 amino acid sequence may be made at the option of the applicant (see Annex 2, Examples 4 and 5).

20. A partial sequence--made up of one or more non-contiguous segments of a larger sequence or of segments from different sequences--should be numbered as a separate sequence, with a separate sequence identifier. A sequence with a gap or gaps should 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. (See Annex 2, Examples 13 and 15).

OTHER AVAILABLE INFORMATION IN THE SEQUENCE LISTING

21. It is recommended that the order and presentation of the items of information in the Sequence Listing follow the order in which those items are listed herein with either the appropriate Data Element Headings or the equivalent Numeric Identifiers. The list of Data Element Headings and equivalent Numeric Identifiers appears in Annex 1 to this Recommendation. When the Data Element Headings are used, they should be in upper case characters, not including any parenthetical explanatory information. When more than one line is necessary for the text following a heading, indentation of the additional line(s) is recommended for use to distinguish it from the heading at the left margin. When a Sequence Listing is prepared using the Numeric Identifiers, the provided information may follow immediately after the Numeric Identifier while only those Numeric Identifiers for which information is given need appear on the Sequence Listing. Generally, a blank line should be inserted between Numeric Identifiers when the digit in the 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 should precede any repeated Numeric Identifier.

WIPO FOR OFFICIAL USE ONLY

22. The Sequence Listing should include, in addition to and immediately preceding the actual nucleotide and/or amino acid sequence, the following items of information, if applicable and when available to the applicant.

GENERAL INFORMATION (Applicant, diskette and application information):

APPLICANT (may be repeated):
NAME (FAMILY NAME, GIVEN NAME(S) and/or INITIAL(S) of a natural person; FULL OFFICIAL DESIGNATION of a legal entity):
STREET (correspondence street address, as elsewhere in application):
CITY (correspondence city address, as elsewhere in application):
STATE OR PROVINCE (correspondence state, as elsewhere in application):
COUNTRY (correspondence country, as elsewhere in application):
POSTAL CODE (correspondence postal code, as elsewhere in application):
TELEPHONE (telephone number of applicant):
TELEFAX (telefax number of applicant):
TELEX OR ELECTRONIC MAIL (telex number or electronic mail address of applicant):

TITLE OF INVENTION (title of the invention, as elsewhere in application):

NUMBER OF SEQUENCES (number of sequences in the “Sequence Listing”):

CORRESPONDENCE ADDRESS:
ADDRESSEE (name of applicant, firm, company or institution, as may be appropriate):
STREET (correspondence street address, as elsewhere in application):
CITY (correspondence city address, as elsewhere in application):
STATE OR PROVINCE (correspondence state, as elsewhere in application):
COUNTRY (correspondence country, as elsewhere in application):
POSTAL CODE (correspondence postal code, as elsewhere in application):

COMPUTER READABLE FORM:
MEDIUM TYPE (type of diskette submitted):
COMPUTER (type of computer used with diskette submitted):
OPERATING SYSTEM (type of operating system used):
SOFTWARE (type of software used to create computer readable form):

CURRENT APPLICATION DATA (if available at time of preparation):
APPLICATION NUMBER:
FILING DATE:
CLASSIFICATION (IPC and/or National):

PRIOR APPLICATION DATA (prior domestic, foreign priority or international application date, if applicable):
APPLICATION NUMBER:
FILING DATE:
CLASSIFICATION (IPC and/or National):

ATTORNEY/AGENT INFORMATION:
NAME (attorney/agent name):
REGISTRATION NUMBER (attorney/agent registration number):
REFERENCE/DOCKET NUMBER (attorney/agent reference or docket number):

TELECOMMUNICATION INFORMATION:
TELEPHONE (telephone number of attorney/agent):
TELEFAX (telefax number of attorney/agent):
TELEX OR ELECTRONIC MAIL (telex number or electronic mail address of attorney/agent):

INFORMATION FOR SEQ ID NO:X:

SEQUENCE CHARACTERISTICS:

LENGTH (sequence length, expressed as number of base pairs or amino acid residues):

TYPE (sequence type, i.e., whether nucleotide or amino acid):

STRANDEDNESS (if nucleic acid, number of strands of source organism molecule, i.e., whether single stranded, double stranded, both, or unknown to applicant):

TOPOLOGY (whether source organism molecule is circular, linear, both, or unknown to applicant):

WIPO FOR OFFICIAL USE ONLY

MOLECULE TYPE (type of molecule sequenced in SEQ ID NO:X)
(at least one of the following should be included with subheadings in Sequence Listing):

- Genomic RNA - Genomic DNA
- mRNA
- tRNA
- rRNA
- snRNA
- scRNA
- preRNA
- cDNA to genomic RNA
- cDNA to mRNA
- cDNA to tRNA
- cDNA to rRNA
- cDNA to snRNA
- cDNA to scRNA
- Other nucleic acid (specify):
- Protein
- Peptide

HYPOTHETICAL (yes/no):

ANTI-SENSE (yes/no):

FRAGMENT TYPE (for proteins and peptides only, at least one of the following should be included in the Sequence Listing):

- N-terminal fragment
- C-terminal fragment
- internal fragment

ORIGINAL SOURCE (original source of molecule sequenced in SEQ ID NO:X):

ORGANISM (scientific name of source organism):

STRAIN:

INDIVIDUAL ISOLATE (name/number of individual isolate):

DEVELOPMENTAL STAGE (give developmental stage of source organism and indicate whether derived from germ-line or rearranged developmental pattern):

HAPLOTYPE:

TISSUE TYPE:

CELL TYPE:

CELL LINE:

ORGANELLE:

IMMEDIATE SOURCE (immediate experimental source of the sequence in SEQ`ID`NO:X):

LIBRARY (library-type, name):

CLONE:

POSITION IN GENOME (position of sequence in SEQ ID NO:X in genome):

WIPO FOR OFFICIAL USE ONLY

CHROMOSOME/SEGMENT (chromosome/segment-name/number):

MAP POSITION:

UNITS (units for map position, i.e., whether units are genome percent, nucleotide number or other (specify)):

FEATURE (description of points of biological significance in the sequence in SEQ ID NO:X) (may be repeated depending on the number of features indicated):

Following the controlled vocabulary of the features field in the existing non-patent literature databases, significant features might include:

active-site
allele
attenuator
binding-site
CAAT signal
cellular
cleavage-site
coding sequence
conflict
cross-links
D-loop
disulfide-bond
domain
duplication
enhancer
exon
GC signal
iDNA
inhibitory-site
insertion sequence
intron LTR (long terminal repeat)
mature peptide
modified base or amino acid
mRNA
mutation
peptide
polyA signal
polyA site
precursor RNA
primary transcript
primer binding
promoter provirus
RBS (ribosome binding site)
repeating unit
repeat region
replication origin
rRNA
satellite
scRNA
signal peptide
snRNA
stem loop
TATA signal
terminator
thiolester-bond
transit peptide
transposon
tRNA
unsure
variation
virion
3’clip
3’UTR

WIPO FOR OFFICIAL USE ONLY

5’clip
5’UTR
-10 signal
-35 signal
or others:

It is recommended to describe features of sequences in accordance with the “The DDBJ/EMBL/GenBank Feature Table: Definition,” obtainable from: DNA Data Bank of Japan, Laboratory of Genetic Information Analysis, Center for Genetic Information Research, National Institute of Genetics, Mishima, Shizuoka 411 Japan; The`European Molecular Biology Laboratory, Postfach 10.2209, 6900 Heidelberg, Germany; NCBI/GenBank, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

NAME/KEY (provide appropriate identifier for feature):

LOCATION:

- 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

IDENTIFICATION METHOD (by which the feature was identified):

- experimentally
- by similarity with a known sequence or to an established consensus sequence
- by similarity to some other pattern

OTHER INFORMATION:

- associated phenotype(s)
- biological/enzymatic activity
- biological/enzymatic activity of its product
- general functional class of the gene and/or gene product
- binding macromolecules--macromolecules to which the gene product can bind
- subcellular localization--subcellular localization of the gene product
- any other relevant information

PUBLICATION INFORMATION (Repeat section for each relevant publication):

AUTHORS:
TITLE (title of publication):
JOURNAL (journal name in which data published):
VOLUME (journal volume in which data published):
ISSUE (journal issue number in which data published):
PAGES (journal page numbers in which data published):
DATE (journal date in which data published):
DOCUMENT NUMBER (document number, for patent type citations only; specify
    two-letter code, number, kind-of-document code):
FILING DATE (document filing date, for patent-type citations only):
PUBLICATION DATE (document publication date; for patent-type citations only):
RELEVANT RESIDUES IN SEQ ID NO:X: FROM TO

WIPO FOR OFFICIAL USE ONLY

SEQUENCE DESCRIPTION: SEQ ID NO:

EXAMPLES

23. To illustrate presentation of the especially important parts of a Sequence Listing, Annexes 2 and 3 are provided. Examples of various kinds of sequence presentation are given in Annex 2 and show presentation not complying with this Recommendation as well as presentation that does comply with this Recommendation. Examples 4, 6, 8, 9, 13, 15 represent properly formatted complete sequences. Examples 5, 10, 11 represent particular excerpts of properly formatted sequences to demonstrate application of parts of the Recommendation. Examples of features indication in a Sequence Listing are given in Annex 3.

24. The complete presentation of data in a Sequence Listing is shown in Annex`4 which gives a specimen Sequence Listing illustrating the major points of this Recommendation.

IMPLEMENTATION OF THIS RECOMMENDATION

25. Each office or other entity choosing to adopt this Recommendation should publish an announcement to that effect in its official journal or gazette. Such an announcement should include an identification of the Sequence Listing Data Elements which the office holds to be mandatory, whether the office will accept a Sequence Listing in Data Element Heading form, in Numeric Identifier form or in either form and, if not previously published, the language(s) to be used.

[Annexes follow]

WIPO FOR OFFICIAL USE ONLY

ANNEX 1

DATA ELEMENT HEADINGS AND NUMERIC IDENTIFIERS

Data Element Heading	Numeric Identifier
GENERAL INFORMATION:	<100>
APPLICANT:	<110>
NAME:	<111>
STREET:	<112>
CITY:	<113>
STATE OR PROVINCE:	<114>
COUNTRY:	<115>
POSTAL CODE:	<116>
TELEPHONE :	<117>
TELEFAX:	<118>
TELEX OR ELECTRONIC MAIL:	<119>
TITLE OF INVENTION:	<120>
NUMBER OF SEQUENCES:	<130>
CORRESPONDENCE ADDRESS:	<140>
ADDRESSEE:	<141>
STREET:	<142>
CITY:	<143>
STATE OR PROVINCE:	<144>
COUNTRY:	<145>
POSTAL CODE:	<146>
COMPUTER READABLE FORM:	<150>
MEDIUM TYPE:	<151>
COMPUTER:	<152>
OPERATING SYSTEM:	<153>
SOFTWARE:	<154>
CURRENT APPLICATION DATA:	<160>
APPLICATION NUMBER:	<161>
FILING DATE:	<162>
CLASSIFICATION:	<163>
PRIOR APPLICATION DATA:	<170>
APPLICATION NUMBER:	<171>
FILING DATE:	<172>
CLASSIFICATION:	<173>
ATTORNEY/AGENT INFORMATION:	<180>
NAME:	<181>
REGISTRATION NUMBER:	<182>
REFERENCE/DOCKET NUMBER:	<183>
TELECOMMUNICATION INFORMATION:	<190>
TELEPHONE :	<191>
TELEFAX:	<192>
TELEX OR ELECTRONIC MAIL:	<193>

WIPO FOR OFFICIAL USE ONLY

Annex 1, page 2

Data Element Heading	Numeric Identifier
INFORMATION FOR SEQ ID NO: X:	<200>
SEQUENCE CHARACTERISTICS:	<210>
LENGTH:	<211>
TYPE:	<212>
STRANDEDNESS:	<213>
TOPOLOGY:	<214>
MOLECULE TYPE:	<220>
HYPOTHETICAL:	<230>
ANTI-SENSE:	<240>
FRAGMENT TYPE:	<250>
ORIGINAL SOURCE:	<260>
ORGANISM:	<261>
STRAIN:	<262>
INDIVIDUAL ISOLATE:	<263>
DEVELOPMENTAL STAGE:	<264>
HAPLOTYPE :	<265>
TISSUE TYPE:	<266>
CELL TYPE:	<267>
CELL LINE:	<268>
ORGANELLE:	<269>
IMMEDIATE SOURCE:	<270>
LIBRARY:	<271>
CLONE:	<272>
POSITION IN GENOME:	<280>
CHROMOSOME/SEGMENT:	<281>
MAP POSITION:	<282>
UNITS:	<283>
FEATURE:	<290>
NAME/KEY:	<291>
LOCATION:	<292>
IDENTIFICATION METHOD:	<293>
OTHER INFORMATION:	<294>
PUBLICATION INFORMATION:	<300>
AUTHORS:	<301>
TITLE:	<302>
JOURNAL:	<303>
VOLUME:	<304>
ISSUE:	<305>
PAGES:	<306>
DATE:	<307>
DOCUMENT NUMBER:	<308>
FILING DATE:	<309>
PUBLICATION DATE:	<310>
RELEVANT RESIDUES IN SEQ ID NO:X: FROM TO	<311>
SEQUENCE DESCRIPTION: SEQ ID NO:	<400>

[Annex 2 follows]

WIPO FOR OFFICIAL USE ONLY

ANNEX 2

EXAMPLES OF SEQUENCE INFORMATION

PRESENTATION OF AMINO ACID SEQUENCES

1. Unformatted Presentation of the Peptide Sequence in Single Letter Code (not according to Recommendation)
2. Unformatted Presentation of the Same Peptide Sequence in Three Letter Code (not according to Recommendation)
3. Unformatted Separated Presentation of the Same Peptide Sequence in Three Letter Code (not according to Recommendation)
4. Formatted Presentation of the Same Peptide Sequence (according to Recommendation)
5. Formatted Peptide Sequence Showing Mature Protein Numbering Scheme (according to Recommendation)
6. Formatted Peptide Sequence with Modified or Unusual Amino Acids at Positions 4, 14, 37 and 38 (according to Recommendation)

PRESENTATION OF NUCLEOTIDE SEQUENCES

7. Unformatted Presentation of the Nucleotide Sequence (not according to Recommendation)
8. Formatted Presentation of the Same Nucleotide Sequence (according to Recommendation)
9. Formatted Nucleotide Sequence with Coding Region and the Peptide Coded by that Sequence (according to Recommendation)
10. Format for Transition from Noncoding to Coding Region (according to Recommendation)
11. Format for a Split Codon (according to Recommendation)

PRESENTATION OF SEQUENCES OF WHICH ONLY PARTS ARE KNOWN

12. Unformatted Peptide Sequence of Which Only Parts are Known (not according to Recommendation)
13. Formatted Presentation of the Same Peptide Sequence (according to Recommendation)
14. Unformatted Nucleotide Sequence of Which Only Parts are Known (not according to Recommendation)
15. Formatted Presentation of the Same Nucleotide Sequence (according to Recommendation)

WIPO FOR OFFICIAL USE ONLY

Annex 2, page 2

1. UNFORMATTED PRESENTATION OF THE PEPTIDE SEQUENCE IN SINGLE LETTER CODE

(<u>not</u> according to Recommendation)

```
AGLLAGSWIPDWTFVSVPPLVTLWYTLTKEPIPGEDVYYVDGACNRNSREGKAGYITQQG
          KQRVEKLENTINQQAELTAIKMALEDSGPRVNIVTDSQYA
```

2. UNFORMATTED PRESENTATION OF THE SAME PEPTIDE SEQUENCE IN THREE LETTER CODE

(<u>not</u> according to Recommendation)

```
AlaGlyLeuLeuAlaGlySerTrpIleProAspTrpThrPheValSerValProProLeu

ValThrLeuTrpTyrThrLeuThrLysGluProIleProGlyGluAspValTyrTyrVal

AspGlyAlaCysAsnArgAsnSerArgGluGlyLysAlaGlyTyrIleThrGlnGlnGly

LysGlnArgValGluLysLeuGluAsnThrThrAsnGlnGlnAlaGluLeuThrAlaIle

LysMetAlaLeuGluAspSerGlyProArgValAsnIleValThrAspSerGlnTyrAla
```

3. UNFORMATTED SEPARATED PRESENTATION OF THE SAME PEPTIDE SEQUENCE IN THREE LETTER CODE

(<u>not</u> according to Recommendation)

```
Ala Gly Leu Leu Ala Gly Ser Trp Ile Pro Asp Trp Thr Phe Val Ser

Val Pro Pro Leu Val Thr Leu Trp Tyr Thr Leu Thr Lys Glu Pro Ile

Pro Gly Glu Asp Val Tyr Tyr Val Asp Gly Ala Cys Asn Arg Asn Ser

Arg Glu Gly Lys Ala Gly Tyr Ile Thr Gln Gln Gly Lys Gln Arg Val

Glu Lys Leu Glu Asn Thr Thr Asn Gln Gln Ala Glu Leu Thr Ala Ile

Lys Met Ala Leu Glu Asp Ser Gly Pro Arg Val Asn Ile Val Thr Asp

Ser Gln Tyr Ala
```

4. FORMATTED PRESENTATION OF THE SAME PEPTIDE SEQUENCE

(according to Recommendation)

```
Ala Gly Leu Leu Ala Gly Ser Trp Ile Pro Asp Trp Thr Phe Val Ser
 1               5                  10                  15
Val Pro Pro Leu Val Thr Leu Trp Tyr Thr Leu Thr Lys Glu Pro Ile
             20                  25                  30
Pro Gly Glu Asp Val Tyr Tyr Val Asp Gly Ala Cys Asn Arg Asn Ser
         35                  40                  45
Arg Glu Gly Lys Ala Gly Tyr Ile Thr Gln Gln Gly Lys Gln Arg Val
     50                  55                  60
Glu Lys Leu Glu Asn Thr Thr Asn Gln Gln Ala Glu Leu Thr Ala Ile
 65                  70                  75                  80
Lys Met Ala Leu Glu Asp Ser Gly Pro Arg Val Asn Ile Val Thr Asp
                 85                  90                  95
Ser Gln Tyr Ala
            100
```

WIPO FOR OFFICIAL USE ONLY

Annex 2, page 3

5. FORMATTED PEPTIDE SEQUENCE SHOWING MATURE PROTEIN NUMBERING SCHEME (according to Recommendation)

```
Met Asn Arg Gly Val Pro Phe Arg His Leu Leu Leu Val Leu Gln Leu
-50                 -45                 -40                 -35
Ala Leu Leu Pro Ala Ala Thr Gln Gly Lys Lys Val Val Leu Gly Lys
                -30                 -25                 -20
Lys Gly Asp Thr Val Glu Leu Thr Cys Thr Ala Ser Gln Lys Lys Ser
            -15                 -10                  -5
Ile Gln Phe His Trp Lys Asn Ser Asn Gln Ile Lys Ile Leu Gly Asn
         1               5                   10
Gln Gly Ser Phe Leu Thr Lys Gly Pro Ser Lys Leu Asn Asp Arg Ala
 15                  20                  25                  30
Asp Ser Arg Arg Ser Leu Trp Asp Gln Gly Asn Phe Pro Leu Ile Ile
                 35                  40                  45
Lys Asn Leu Lys Ile Glu ... ... ...
             50
```

6. FORMATTED PEPTIDE SEQUENCE WITH MODIFIED OR UNUSUAL AMINO ACIDS AT POSITIONS 4, 14, 37 AND 38 (according to Recommendation)

```
Ala Gly Leu Leu Ala Gly Ser Trp Ile Pro Asp Trp Thr Xaa Val Ser
 1               5                   10                  15
Val Pro Pro Leu Val Thr Leu Trp Tyr Thr Leu Thr Lys Glu Pro Ile
             20                  25                  30
Pro Gly Glu Asp Val Xaa Tyr Val Asp Gly Ala Cys Asn Arg Asn Ser
         35                  40                  45
Arg Glu Gly Lys Ala Gly Tyr Ile Thr Gln Gln Gly Lys Gln Arg Val
     50                  55                  60
Glu Lys Leu Glu Asn Thr Thr Asn Gln Gln Ala Glu Leu Thr Ala Ile
 65                      70              75                  80
Lys Met Ala Leu Glu Asp Ser Gly Pro Arg Val Asn Ile Val Thr Asp
                 85                  90                  95
Ser Gln Tyr Ala
            100
```

To indicate 2-Aminoadipic acid at position 14 and 3-Aminoadipic acid at position 38 using the abbreviations of paragraph 12, a Sequence Listing would contain the following:

FEATURE:

NAME/KEY: misc-feature
LOCATION: 14
OTHER INFORMATION: Xaa is Aad

FEATURE:

NAME/KEY: misc-feature
LOCATION: 38
OTHER INFORMATION: Xaa is bAad

WIPO FOR OFFICIAL USE ONLY

As described in paragraph 12, in some cases unmodified amino acids may represent modified or unusual amino acids in the sequence, e.g., to indicate Norleucine at position`4 and Norvaline at position 37, a Sequence Listing would contain the following:

FEATURE:

NAME/KEY: misc-feature
LOCATION: 4
OTHER INFORMATION: Leu is Nle

FEATURE:

NAME/KEY: misc-feature
LOCATION: 37
OTHER INFORMATION: Val is Nva

7. UNFORMATTED PRESENTATION OF THE NUCLEOTIDE SEQUENCE (<u>not</u> according to Recommendation)

```
CAAGCCCAGAGCCCTGCCATTTCTGTGGGCTCAGGTCCCTACTGCTCAGCCCCTTCCTCC
CTCGGCAAGGCCACAATGAACCGGGGAGTCCCTTTTAGGCACTTGCTTCTGGTGCTGCAA
CTGGCGCTCCTCCCAGCAGCCACTCAGGGAAAGAAAGTGGTGCTGGGCAAAAAAGGGGAT
ACAGTGGAACTGACCTGTACAGCTTCCCAGAAGAAGAGCATACAATTCCACTGGAAAAAC
TCCAACCAGATAAAGATTCTGGGAAATCAGGGCTCCTTCTTAACTAAAGGTCCATCCAAG
CTGAATGATCGCGCTGACTCAAGAAGAAGCCTTTGGGACCAAGGAAACTTCCCCCTGATC
ATCAAGAATCTTAAGATAGAAGACTCAGATACTTACATCTGTGAAGTGGAGGACCAGAAG
GAGGAGGTGCAATTGCTAGTGTTCGGATTGACTGCCAACTCTGACACCCACCTGCTTCAG
GGGCAGAGCCTGACCCTGACCTTGGAGAGCCCCCCTGGTAGTAGCCCCTCAGTGCAATGT
AGGAGTCCAAGGGGTAAAAACATACAGGGGGGGAAGACCCTCTCCGTGTCTCAGCTGGAG
CTCCAGGATAGTGGCACCTGGACATGCACTGTCTTGCAGAACCAGAAGAAGGTGGAGTTC
AAAATAGACATCGTGGTGCTAGCTTTCCAGAAGGCCTCCAGCATAGTCTATAAGAAAGAG
GGGGAACAGGTGGAGTTCTCCTTCCCACTCGCCTTTACAGTTGAAAAGCTGACGGGCAGT
GGCGAGCTGTGGTGGCAGGCGGAGAGGGCTTCCTCCTCCAAGTCTTGGATCACCTTTGAC
CTGAAGAACAAGGAAGTGTCTGTAAAACGGGTTACCCAGGACCCTAAGCTCCAGATGGGC
AAGAAGCTCCCGCTCCACCTCACCCTGCCCCAGGCCTTGCCTCAGTATGCTGGCTCTGGA
AACCTCACCCTGGCCCTTGAAGCGAAAACAGGAAAGTTGCATCAGGAAGTGAACCTGGTG
GTGATGAGAGCCACTCAGCTCCAGAAAAATTTGACCTGTGAGGTGTGGGGACCCACCTCC
CCTAAGCTGATGCTGAGCTTGAAACTGGAGAACAAGGAGGCAAAGGTCTCGAAGCGGGAG
AAGGCGGTGTGGGTGCTGAACCCTGAGGCGGGGATGTGGCAGTGTCTGCTGAGTGACTCG
GGACAGGTCCTGCTGGAATCCAACATCAAGGTTCTGCCCACATGGTCCACCCCGGTGCAG
CCAATGGCCCTGATTGTGCTGGGGGGCGTCGCCGGCCTCCTGCTTTTCATTGGGCTAGGC
ATCTTCTTCTGTGTCAGGTGCCGGCACCGAAGGCGCCAAGCAGAGCGGATGTCTCAGATC
AAGAGACTCCTCAGTGAGAAGAAGACCTGCCAGTGCCCTCACCGGTTTCAGAAGACATGT
AGCCCCATTTGAGGCACGAGGCCAGGCAGATCCCACTTGCAGCCTCCCCAGGTGTCTGCC
CCGCGTTTCCTGCCTGCGGACCAGATGAATGTAGCAGATCCCACGCTCTGGCCTCCTGTT
CGTCCTCCCTACAATTTGCCATTGTTTCTCCTGGGTTAGGCCCCGGCTTCACTGGTTGAG
TGTTGCTCTCTAGTTTCCAGAGGCTTAATCACACCGTCCTCCACGCCATTTCCTTTTCCT
TCAAGCCTAGCCCTTCTCTCATTATTTCTCTCTGACCCTCTCCCCACTGCTCATTTGGAT
CC
```

WIPO FOR OFFICIAL USE ONLY

Annex 2, page 5

8. FORMATTED PRESENTATION OF THE SAME NUCLEOTIDE SEQUENCE
(according to Recommendation)

```
CAAGCCCAGA  GCCCTGCCAT  TTCTGTGGGC  TCAGGTCCCT  ACTGCTCAGC  CCCTTCCTCC     60
CTCGGCAAGG  CCACAATGAA  CCGGGGAGTC  CCTTTTAGGC  ACTTGCTTCT  GGTGCTGCAA    120
CTGGCGCTCC  TCCCAGCAGC  CACTCAGGGA  AAGAAAGTGG  TGCTGGGCAA  AAAAGGGGAT    180
ACAGTGGAAC  TGACCTGTAC  AGCTTCCCAG  AAGAAGAGCA  TACAATTCCA  CTGGAAAAAC    240
TCCAACCAGA  TAAAGATTCT  GGGAAATCAG  GGCTCCTTCT  TAACTAAAGG  TCCATCCAAG    300
CTGAATGATC  GCGCTGACTC  AAGAAGAAGC  CTTTGGGACC  AAGGAAACTT  CCCCCTGATC    360
ATCAAGAATC  TTAAGATAGA  AGACTCAGAT  ACTTACATCT  GTGAAGTGGA  GGACCAGAAG    420
GAGGAGGTGC  AATTGCTAGT  GTTCGGATTG  ACTGCCAACT  CTGACACCCA  CCTGCTTCAG    480
GGGCAGAGCC  TGACCCTGAC  CTTGGAGAGC  CCCCCTGGTA  GTAGCCCCTC  AGTGCAATGT    540
AGGAGTCCAA  GGGGTAAAAA  CATACAGGGG  GGGAAGACCC  TCTCCGTGTC  TCAGCTGGAG    600
CTCCAGGATA  GTGGCACCTG  GACATGCACT  GTCTTGCAGA  ACCAGAAGAA  GGTGGAGTTC    660
AAAATAGACA  TCGTGGTGCT  AGCTTTCCAG  AAGGCCTCCA  GCATAGTCTA  TAAGAAAGAG    720
GGGGAACAGG  TGGAGTTCTC  CTTCCCACTC  GCCTTTACAG  TTGAAAAGCT  GACGGGCAGT    780
GGCGAGCTGT  GGTGGCAGGC  GGAGAGGGCT  TCCTCCTCCA  AGTCTTGGAT  CACCTTTGAC    840
CTGAAGAACA  AGGAAGTGTC  TGTAAAACGG  GTTACCCAGG  ACCCTAAGCT  CCAGATGGGC    900
AAGAAGCTCC  CGCTCCACCT  CACCCTGCCC  CAGGCCTTGC  CTCAGTATGC  TGGCTCTGGA    960
AACCTCACCC  TGGCCCTTGA  AGCGAAAACA  GGAAAGTTGC  ATCAGGAAGT  GAACCTGGTG   1020
GTGATGAGAG  CCACTCAGCT  CCAGAAAAAT  TTGACCTGTG  AGGTGTGGGG  ACCCACCTCC   1080
CCTAAGCTGA  TGCTGAGCTT  GAAACTGGAG  AACAAGGAGG  CAAAGGTCTC  GAAGCGGGAG   1140
AAGGCGGTGT  GGGTGCTGAA  CCCTGAGGCG  GGGATGTGGC  AGTGTCTGCT  GAGTGACTCG   1200
GGACAGGTCC  TGCTGGAATC  CAACATCAAG  GTTCTGCCCA  CATGGTCCAC  CCCGGTGCAG   1260
CCAATGGCCC  TGATTGTGCT  GGGGGGCGTC  GCCGGCCTCC  TGCTTTTCAT  TGGGCTAGGC   1320
ATCTTCTTCT  GTGTCAGGTG  CCGGCACCGA  AGGCGCCAAG  CAGAGCGGAT  GTCTCAGATC   1380
AAGAGACTCC  TCAGTGAGAA  GAAGACCTGC  CAGTGCCCTC  ACCGGTTTCA  GAAGACATGT   1440
AGCCCCATTT  GAGGCACGAG  GCCAGGCAGA  TCCCACTTGC  AGCCTCCCCA  GGTGTCTGCC   1500
CCGCGTTTCC  TGCCTGCGGA  CCAGATGAAT  GTAGCAGATC  CCACGCTCTG  GCCTCCTGTT   1560
CGTCCTCCCT  ACAATTTGCC  ATTGTTTCTC  CTGGGTTAGG  CCCCGGCTTC  ACTGGTTGAG   1620
TGTTGCTCTC  TAGTTTCCAG  AGGCTTAATC  ACACCGTCCT  CCACGCCATT  TCCTTTTCCT   1680
TCAAGCCTAG  CCCTTCTCTC  ATTATTTCTC  TCTGACCCTC  TCCCCACTGC  TCATTTGGAT   1740
CC                                                                    1742
```

9. FORMATTED NUCLEOTIDE SEQUENCE WITH CODING REGION AND THE PEPTIDE CODED BY THAT SEQUENCE
(according to Recommendation)

```
CAAGCCCAGA GCCCTGCCAT TTCTGTGGGC TCAGGTCCCT ACTGCTCAGC CCCTTCCTCC         60

 CTCGGCAAGG CCACA ATG AAC CGG GGA GTC CCT TTT AGG CAC TTG CTT CTG       111
                  Met Asn Arg Gly Val Pro Phe Arg His Leu Leu Leu
                                  -45                 -40
 GTG CTG CAA CTG GCG CTC CTC CCA GCA GCC ACT CAG GGA AAG AAA GTG       159
 Val Leu Gln Leu Ala Leu Leu Pro Ala Ala Thr Gln Gly Lys Lys Val
         -35                 -30                 -25
 GTG CTG GGC AAA AAA GGG GAT ACA GTG GAA CTG ACC TGT ACA GCT TCC       207
 Val Leu Gly Lys Lys Gly Asp Thr Val Glu Leu Thr Cys Thr Ala Ser
     -20                 -15                 -10
 CAG AAG AAG AGC ATA CAA TTC CAC TGG AAA AAC TCC AAC CAG ATA AAG       255
 Gln Lys Lys Ser Ile Gln Phe His Trp Lys Asn Ser Asn Gln Ile Lys
  -5                   1               5                   10
 ATT CTG GGA AAT CAG GGC TCC TTC TTA ACT AAA GGT CCA TCC AAG CTG       303
 Ile Leu Gly Asn Gln Gly Ser Phe Leu Thr Lys Gly Pro Ser Lys Leu
              15                  20                  25
```

WIPO FOR OFFICIAL USE ONLY

9. FORMATTED NUCLEOTIDE SEQUENCE WITH CODING REGION AND THE PEPTIDE CODED BY THAT SEQUENCE (cont'd)

```
AAT GAT CGC GCT GAC TCA AGA AGA AGC CTT TGG GAC CAA  GGA AAC TTC              351
Asn Asp Arg Ala Asp Ser Arg Arg Ser Leu Trp Asp Gln  Gly Asn Phe
        30                  35                  40
CCC CTG ATC ATC AAG AAT CTT AAG ATA GAA GAC TCA GAT  ACT TAC ATC              399
Pro Leu Ile Ile Lys Asn Leu Lys Ile Glu Asp Ser Asp  Thr Tyr Ile
    45                  50                  55
TGT GAA GTG GAG GAC CAG AAG GAG GAG GTG CAA TTG CTA  GTG TTC GGA              447
Cys Glu Val Glu Asp Gln Lys Glu Glu Val Gln Leu Leu  Val Phe Gly
60                  65                  70                   75
TTG ACT GCC AAC TCT GAC ACC CAC CTG CTT CAG GGG CAG  AGC CTG ACC              495
Leu Thr Ala Asn Ser Asp Thr His Leu Leu Gln Gly Gln  Ser Leu Thr
                80                  85                   90
CTG ACC TTG GAG AGC CCC CCT GGT AGT AGC CCC TCA GTG  CAA TGT ATT              543
Leu Thr Leu Glu Ser Pro Pro Gly Ser Ser Pro Ser Val  Gln Cys Ile
            95                  100                  105
ATG GCC CTG ATT GTG CTG GGG GGC GTC GCC GGC CTC CTG  CTT TTC ATT              591
Met Ala Leu Ile Val Leu Gly Gly Val Ala Gly Leu Leu  Leu Phe Ile
        110                 115                 120
GGG CTA GGC ATC TTC TTC TGT GTC AGG TGC CGG CAC CGA  AGG CGC CAA              639
Gly Leu Gly Ile Phe Phe Cys Val Arg Cys Arg Cys His  Arg Arg Gln
    125                 130                 135
GCA GAG CGG ATG TCT CAG ATC AAG AGA CTC CTC AGT GAG  AAG AAG ACC              687
Ala Glu Arg Met Ser Gln Ile Lys Arg Leu Leu Ser Glu  Lys Lys Thr
140                 145                 150
TGC CAG TGC CCT CAC CGG TTT CAG AAG ACA TGT AGC CCC  ATT TGAGGCACGA           739
Cys Gln Cys Pro His Arg Phe Gln Lys Thr Cys Ser Pro  Ile
                160                 165

GGCCAGGCAG ATCCCACTTG CAGCCTCCCC AGGTGTCTGC CCCGCGTTTC CTGCCTGCGG            799

ACCAGATGAA TGTAGCAGAT CCCACGCTCT GGCCTCCTGT TCGTCCTCCC TACAATTTGC            859

CATTGTTTCT CCTGGGTTAG GCCCCGGCTT CACTGGTTGA GTGTTGCTCT CTAGTTTCCA            919
```

I0. FORMAT FOR TRANSITION FROM NONCODING TO CODING REGION (according to Recommendation)

```
CAAGCCCAGA GCCCTGCCAT TTCTGTGGGC TCAGGTCCCT ACTGCTCAGC CCCTTCCTCC             345

CTCGGCAAGG CCACA ATG AAC CGG GGA GTC CCT TTT AGG CAC TTG CTT CTG              111
                 Met Asn Arg Gly Val Pro Phe Arg His Leu Leu Leu
                                 -45                 -40
```

Annex 2, page 7

11. FORMAT FOR A SPLIT CODON
(according to Recommendation)

```
........... GGT AAA AAC ATA CAG GGG GGG AAG ACC C  TCAGACTCCA           368

........... Gly Lys Asn Ile Gln Gly Gly Lys Thr
                                115

TCTGTCTGT  TC CAG  GAT  AGT  CGC  ACC  TGG  ACA  TGC  ACT ...........

Leu Gln Asp Ser Gly Thr Trp Thr Cys Thr   ...........
120                 125
```

CTC codes for the Leu at 120. The Leu in the peptide sequence is positioned under the group of two nucleotide bases TC rather than being split or positioned under the C at 358 in the nucleotide chain.

12. UNFORMATTED PEPTIDE SEQUENCE OF WHICH ONLY PARTS ARE KNOWN
(not according to Recommendation)

```
Ala Gly Leu Leu Ala Gly Ser Trp Ile Pro Asp Trp Thr Phe Val Ser
Val Pro Pro Leu Val Thr Leu Trp .......................... Ile
Pro Gly Glu Asp Val Tyr Tyr Val Asp Gly Ala Cys Asn Arg Asn Ser
Arg Glu Gly .............................. Lys Gln Arg Val Glu
Lys Leu Glu Asn Thr Thr Asn Gln Gln Ala Glu Leu Thr Ala Ile Lys
Met Ala Leu Glu Asp Ser Gly Pro Arg Val Asn Ile Val Thr Asp Ser
Gln Tyr Ala
```

Unknown regions are represented by ....

13. FORMATTED PRESENTATION OF THE SAME PEPTIDE SEQUENCE
(according to Recommendation)

SEQ ID NO:1:

```
Ala Gly Leu Leu Ala Gly Ser Trp Ile Pro Asp Trp Thr Phe Val Ser
1               5                   10                  15
Val Pro Pro Leu Val Thr Leu Trp
            20
```

SEQ ID NO:2:

```
Ile Pro Gly Glu Asp Val Tyr Tyr Val Asp Gly Ala Cys Asn Arg Asn
1               5                   10                  15
Ser Arg Glu Gly
            20
```

SEQ ID NO:3:

```
Lys Gln Arg Val Glu Lys Leu Glu Asn Thr Thr Asn Gln Gln Ala Glu
1               5                   10                  15
Leu Thr Ala Ile Lys Met Ala Leu Glu Asp Ser Gly Pro Arg Val Asn
            20                  25                  30
Ile Val Thr Asp Ser Gln Tyr Ala
        35                  40
```

WIPO FOR OFFICIAL USE ONLY

Annex 2, page 8

14. UNFORMATTED NUCLEOTIDE SEQUENCE OF WHICH ONLY PARTS ARE KNOWN
(<u>not</u> according to Recommendation)

```
AAATAGACCTCACCCTTACCCACTTCCCCTAGCGCTGAAA....................
............AAATTTGCCTAGTCAAAATAAAAGATGCCGAGTCTATAAAAGCGCAAG
GACAGTTCAGGAGGTGGCTCGCTCCCTCACCGACCCTCTGGTCACGGAGACTCACCTTGG
GGATCCATCCTC..........................TCCGTGGGACCGTCTCCCGGCC
TCGGCACCTCCTGAACTGCTCCTCCCAAGGTAAGTCTCCTCTAGGGTCGAGCTCGGCTGC
CCCTTAGGTACTCGCTCCCCGAGGGTCTTTAGAGACAC......................
CTAGACTCTGCCTTAAACTTCACTTCCGCGTTCTTGTCTCGTTCTTTCCTCTTCGCCGTC
ACTGAAAACGAAACCTCAACGCCGCCCTCTTGGC..........................
TGGAGCGCAGCAAGGGCTAGGGCTTCCTGAACCTCTCCGGGAGAGGTCTATTGCTATAGG
CAGGCCCGCCCTAGGAGCATTGTCTTCCCGGGGAAGACAAACAATTGGGGGCTCGTCCGG
GATTTGAATTCCTCCATTCTCACATTATGGGACAAATCCACGGGCTTTCCCCAACTCCAA
```

Unknown regions are indicated by ....

15. FORMATTED PRESENTATION OF THE SAME NUCLEOTIDE SEQUENCE
(according to Recommendation)

SEQ ID NO:1

```
AAATAGACCT CACCCTTACC CACTTCCCCT AGCGCTGAAA                                40
```

SEQ ID NO:2

```
AAATTTGCCT AGTCAAAATA AAAGATGCCG AGTCTATAAA AGCGCAAGGA CAGTTCAGGA          60
GGTGGCTCGC TCCCTCACCG ACCCTCTGGT CACGGAGACT CACCTTGGGG ATCCATCCTC         120
```

SEQ ID NO:3

```
TCCGTGGGAC CGTCTCCCGG CCTCGGCACC TCCTGAACTG CTCCTCCCAA GGTAAGTCTC          60
CTCTCAGGTC GAGCTCGGCT GCCCCTTAGG TAGTCGCTCC CCGAGGGTCT TTAGAGACAC         120
```

SEQ ID NO:4 :

```
CTAGACTCTG CCTTAAACTT CACTTCCGCG TTCTTGTCTC GTTCTTTCCT CTTCGCCGTC          60
ACTGAAAACG AAACCTCAAC GCCGCCCTCT TGGC                                      94
```

SEQ ID NO:5 :

```
TGGAGCGCAG CAAGGGCTAG GGCTTCCTGA ACCTCTCCGG GAGAGGTCTA TTGCTATAGG          60
CAGGCCCGCC CTAGGAGCAT TGTCTTCCCG GGGAAGACAA ACAATTGGGG GCTCGTCCGG         120
GATTTGAATT CCTCCATTCT CACATTATGG GACAAATCCA CGGGCTTTCC CCAACTCCAA         180
```

[Annex 3 follows]

WIPO FOR OFFICIAL USE ONLY

ANNEX 3

EXAMPLES OF FEATURES INDICATION IN A SEQUENCE LISTING

FEATURE:

NAME/KEY: exon

LOCATION: 194..264

OTHER INFORMATION:
/product=”Thyrotropin beta subunit precursor”

FEATURE:

NAME/KEY: signal peptide

LOCATION: 4213..4272

OTHER INFORMATION:
/product=”Thyrotropin beta subunit signal peptide”
/standard name=”TSH-beta”

FEATURE:

NAME/KEY: mature peptide

LOCATION: join (4273..4374, 4752..5003)

OTHER INFORMATION:
/product=”Thyrotropin beta subunit”
/standard name=”TSH-beta”

FEATURE:

NAME/KEY: coding sequence

LOCATION: join (4213..4374, 4752..5006)

OTHER INFORMATION:
/product=”Thyrotropin beta subunit precursor”
/standard name=”TSH-beta”

[Annex 4 follows]

WIPO FOR OFFICIAL USE ONLY

ANNEX 4

SPECIMEN SEQUENCE LISTING
SEQUENCE LISTING

GENERAL INFORMATION:

APPLICANT:
NAME: Applicant Plc
STREET: 124 Main Street
CITY: Smalltown
STATE: South State
COUNTRY: Newland
POSTAL CODE: 2288 ST
TELEPHONE: 001-1234567
TELEFAX: 001-1234567
TELEX: 12345 apc

NAME: Anmelder GmbH
STREET: Lindenstrasse 27
CITY: Neustadt
COUNTRY: Germany
POSTAL CODE: 12345
TELEPHONE: 002-24680
TELEFAX: 002-24680
TELEX: 2468 agh

TITLE OF INVENTION: Example of a Sequence Listing

NUMBER OF SEQUENCES: 1

COMPUTER READABLE FORM:
MEDIUM TYPE: Floppy disk
COMPUTER: IBM PC compatible
OPERATING SYSTEM: PC-DOS/MS-DOS
SOFTWARE: PatentIn Release #1.0, Version #1.25

CURRENT APPLICATION DATA:
APPLICATION NUMBER: EP 91000000

PRIOR APPLICATION DATA:
APPLICATION NUMBER: US 07/123456
FILING DATE: 30-AUG-1991

INFORMATION FOR SEQ ID NO:1

SEQUENCE CHARACTERISTICS:
LENGTH: 2654 base pairs
TYPE: nucleic acid
STRANDEDNESS: single
TOPOLOGY: linear

MOLECULE TYPE: DNA (genomic)

ORIGINAL SOURCE:
ORGANISM: Homo sapiens
CELL TYPE: Leukocyte

FEATURE:
NAME/KEY: TATA-signal
LOCATION: 339..344

WIPO FOR OFFICIAL USE ONLY

Annex 4, page 2

FEATURE:
NAME/KEY: intron
LOCATION: 527..814

FEATURE:
NAME/KEY: intron
LOCATION: 923..1006

FEATURE:
NAME/KEY: intron
LOCATION: 1113..1359

FEATURE:
NAME/KEY: signal peptide
LOCATION: 824..1009

FEATURE:
NAME/KEY: mature peptide
LOCATION: 1010..1772

FEATURE:
OTHER INFORMATION: human lymphotoxin gene

PUBLICATION INFORMATION:
AUTHORS: Doe, Joan X
Doe, John Q
TITLE: Isolation and Characterization of a Gene Encoding a Protease from Paramecium sp.
JOURNAL: Fictional Genes
VOLUME: 1
ISSUE: 1
PAGES: 1-20
DATE: 2-MAR-1988
RELEVANT RESIDUES IN SEQ ID NO: 1: FROM 1 TO 957

WIPO FOR OFFICIAL USE ONLY

SEQUENCE DESCRIPTION: SEQ ID NO:1 :

```
AAGGGTGCAG AGATGTTATA TATGATTGCT CTTCAGGGAA CCGGCCTCCA GCTCACACCC       60
CAGCTGCTCA ACCGCCTCCT CTCTGAATTG ACTGTCCCTT CTTTGGAACT CTAGGCCTGA      120
CCCCACTCCC TGGCCCTCCC AGCCCACGAT TCCCCTGACC CGACTCCCTT TCCAGGAACT      180
CAGTCGCCTG AACCCCCAGC CTGTGGTTCT CTCCTAGGCC TCAGCCTTTC CTGCCTTTGA      240
CTGAAACAGC AGTATCTTCT ACACGCTGGG GCTTCCCGCG GCCCAGCCCC GACCTAGAAC      300
CCGCCCGCTG CCTGCCACGC TGCCACTGCC GCTTCCTCTA TAAAGGGACC TGAGCGTCCG      360
CGCGCAGGGG CTCCACACAG CAGGTGAGGC TCTCCTGCCC CATCTCCTTG GGCTGCCCGT      420
GCTTCGTGCT TTGGACTACC GCCCCGAGTG TCCTGCCCTC TGCCTGGGCC TCGGTCCCTC      480
CTGCACCTGC TGCCTGGATC CCCGGCCTGC CTGGGCCTGG GCCTTGGTGG GTTTGGTTTT      540
GGTTTCCTTC TCTGTCTCTG ACTCTCCATC TGTCAGTCTC ATTGTCTCTG TCACACATTC      600
TCTGTTTCTG CCATGGTTCC TCTCTGTTCC CTTCCTGTCT CTCTCTGTCT CCCTCTGCTC      660
ACCTTGGGGT TTCTCTGACT GCATCTTGTC CCCTTCTCTG TCCGATCTCT CTCTCGGGGG      720
TCGGGGGGTG CTGTCTCCCA GGGCGGGAGG TCTGTCTTCC GCCGCGTGCC CCGCCCCGCT      780
CACTGTCTCT CTCTCTCTCT CTCTTTCICT GCAGGTTCTC CCC ATG ACA CCA CCT        835
                                                Met Thr Pro Pro

GAA CGT CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC        883
Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu
-30                 -25                 -20                 -15

CTT CTG GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAG GTGAGGCAGC         932
Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala Gln
                -10                 -5

AGGAGAATGG GGGCTGCTGG GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA      992
CTCTGTTCTC CTAG  GGG CTC CCT GGT GTT GGC CTC ACA CCT TCA GCT GCC      1042
                Gly Leu Pro Gly Val Gly Leu Thr Pro Ser Ala Ala
                     1               5                  10

CAG ACT GCC CGT CAG CAC CCC AAG ATG CAT CTT GCC CAC AGC AAC CTC       1090
Gln Thr Ala Arg Gln His Pro Lys Met His Leu Ala His Ser Asn Leu
            15                  20                  25

AAA CCT GCT GCT CAC CTC ATT G  GTAAACATCC ACCTGACCTC CCAGACATGT       1142
Lys Pro Ala Ala His Leu Ile
         30

CCCCACCAGC TCTCCTCCTA CCCCTGCCTC AGGAACCCAA GCATCCACCC CTCTCCCCCA     1202
ACTTCCCCCA CGCTAAAAAA AACAGAGGGA GCCCACTCCT ATGCCTCCCC CTGCCATCCC     1262
CCAGGAACTC AGTTGTTCAG TGCCCACTTC CTCAGGGATT GAGACCTCTG ATCCAGACCC     1322
CTGATCTCCC ACCCCCATCC CCTATGGCTC TTCCTAG GA  GAC CCC AGC AAG CAG      1376
                                         Gly Asp Pro Ser Lys Gln
                                         35                  40

AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT GCC TTC  CTC CAG GAT      1424
Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg Ala Phe  Leu Gln Asp
                 45                  50                  55

GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC CCC ACC  AGT GGC ATC      1472
Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val Pro Thr  Ser Gly Ile
            60                  65                   70
```

WIPO FOR OFFICIAL USE ONLY

```
TAC TTC GTC TAC TCC CAG GTG GTC TTC TCT GGG AAA GCC  TAC TCT CCC      1520
Tyr Phe Val Tyr Ser Gln Val Val Phe Ser Gly Lys Ala  Tyr Ser Pro
         75                  80                  85

AAG GCC ACC TCC TCC CCA CTC TAC CTG GCC CAT GAG GCT  CAG CTC TTC      1568
Lys Ala Thr Ser Ser Pro Leu Tyr Leu Ala His Glu Val  Gln Leu Phe
     90                  95                 100

TCC TCC CAG TAC CCC TTC CAT GTG CCT CTC CTC AGC TCC  CAG AAG ATG      1616
Ser Ser Gln Tyr Pro Phe His Val Pro Leu Leu Ser Ser  Gln Lys Met
105                 110                 115                  120

GTG TAT CCA GGG CTG CAG GAA CCC TGG CTG CAC TCG ATG  TAC CAC GGG      1664
Val Tyr Pro Gly Leu Gln Glu Pro Trp Leu His Ser Met  Tyr His Gly
                125                 130                  135

GCT GCG TTC CAG CTC ACC CAG GGA GAC CAG CTA TCC ACC  CAC ACA GAT      1712
Ala Ala Phe Gln Leu Thr Gln Gly Asp Gln Leu Ser Thr  His Thr Asp
            140                 145                  150

GGC ATC CCC CAC CTA GTC CTC AGC CCT AGT ACT GTC TTC  TTT GGA GGC      1760
Gly Ile Pro His Leu Val Leu Ser Pro Ser Thr Val Phe  Phe Gly Ala
        155                 160                 165

TTC GCT CTG TAGAACTTGG AAAAATCCAG AAAGAAAAAA TAATTGATTT  C            1810
Phe Ala Leu
    170

AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA GGGGTCGTCA CCACCTCTCC     1870
TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC ACCGGAGCTT TCAAAGAAGG     1930
AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA TCCCTGATGT CTGTCTGGCT     1990
GAGGATTTCA AGCCTGCCTA GGAATTCCCA GCCCAAAGCT GTTGGTCTTG TCCCACCAGC     2050
TAGGTGGGGC CTAGATCCAC ACACAGAGGA AGAGCAGGCA CATGGAGGAG CTTGGGGGAT     2110
GACTAGAGGC AGGGAGGGGA CTATTTATGA AGGCCAAAAA ATTAAATTAT TTATTTATGG     2170
AGGATGGAGA GAGGGGAATA ATAGAAGAAC ATCCAAGGAG AAACAGAGAC AGGCCCAAGA     2230
GATGAAGAGT GAGAGGGCAT GCGCACAAGG CTGACCAAGA GAGAAAGAAG TAGGCATGAG     2290
GGATCACAGG GCCCCAGAAG GCAGGGAAAG GCTCTGAAAG CCAGCTGCCG ACCAGAGCCC     2350
CACACGGAGG CATCTGCACC CTCGATGAAG CCCAATAAAC CTCTTTTCTC TGAAATGCTG     2410
TCTGCTTGTG TGTGTGTGTC TGGGAGTGAG AACTTCCCAG TCTATCTAAG GAATGGAGGG     2470
AGGGACAGAG GGCTCAAAGG GACGAAGAGC TGTGGGGAGA ACAAAAGGAT AAGGGCTCGA     2530
GAGAGCTTCA AGGATATGTG ATGGATCACC AGGTGAGGCC GCCAGACTGC TGCAGGGGAA     2590
GCAAAGGAGA AGCTGAGAAG ATGAAGGAAA AGTCAGGGTC TGGAGGGGGG GGGGTCAGGG     2650
AGCT                                                                  2654
```

[End of Annex and of Standard]

WIPO FOR OFFICIAL USE ONLY