WIPO

PCT/MIA/IV/9
ORIGINAL: English
DATE: June 14, 1994

WORLD INTELLECTUAL PROPERTY ORGANIZATION
GENEVA

INTERNATIONAL PATENT COOPERATION UNION (PCT UNION)

MEETING OF INTERNATIONAL AUTHORITIES UNDER THE PCT

Fourth Session
Geneva, June 27 to July 1, 1994

PROPOSAL CONCERNING THE LANGUAGE OF NUCLEOTIDE AND/OR AMINO ACID SEQUENCE LISTINGS

Document prepared by the International Bureau

The Annex to this document contains a proposal concerning the language of nucleotide and/or amino acid sequence listings disclosed in international applications, submitted for consideration at the fourth session of the Meeting of International Authorities under the PCT. This proposal is based on a proposal that was the subject of an exchange of views among the European Patent Office, the Japanese Patent Office and the United States Patent and Trademark Office in Tokyo in May 1994, during a technical meeting in the context of those Offices' trilateral cooperation.

[Annex follows]

SEQUENCE LISTINGS

I. INTRODUCTION

1. The present document addresses the question of international applications which, in accordance with Rules 5.2 and 13ter PCT, must contain a nucleotide and/or amino-acid sequence listing (SL) on paper and in machine-readable form.

2. Most specifically, the aim of the document is to put under discussion in MIA an outline for a common PCT standard to allow the applicant to draw up a single sequence listing on paper and in machine-readable form which would be acceptable to the competent ISA and to the designated/ elected Offices.

3. The problem here is one of language, as explained below.

II. THE LANGUAGE PROBLEM

4. A SL is a highly specialised technical description, the core of which are the sequences themselves, written down in the universally accepted genetic alphabet (nucleotides) and/or a three-letter code (amino acids).
This part of any SL is language-independent.
In addition, the SL also contains general information (bibliographic data relating to the applicant and the application) and data relating to each sequence (such as length, type, strandedness etc.).

5. The updated WIPO Standard ST. 23 issued in 1993[1] has rationalised the presentation of the general information and other data elements by recommending the use of numeric identifiers for all data element headings. As a result, the data element HEADINGS in any SL are also language-independent (see Annex 1 to ST. 24 = Annex 1 to this paper).

…/…

[1] The revised text was adopted by the PCPI Executive Coordination Committee at its session from 13 to 17 December 1993.

6. As to the data elements per se, four different categories can be distinguished:

(a) language-independent bibliographic data (relating to the applicant etc.);

(b) feature data, relating to sequences, of the kind given in internationally recognised lists of abbreviations and technical terms, and thus regarded as language-independent. It is proposed to use the DDBJ/EMBL/GenBank Feature Table[2], as recommended in WIPO Standard ST. 23, point 22;

(c) language-dependent data elements, relating to the sequences, of the kind not yet covered by the standard and/or the feature table;

(d) language-dependent data terms comprising free text.

Significance of language-dependent features

7. The following considerations need to be borne in mind when assessing the significance of language-dependent data terms in SLs for patent offices and patent information users:

The Trilateral Offices (EPO, JPO, USPTO) have made the following proposal regarding compulsory and optional elements to be included in SLs:

(i) **only** numeric identifiers of data element headings, as defined in ST. 23 and ST. 24, should be used in SLs submitted under the PCT and national/regional procedures;

(ii) not all of the data headings listed in ST. 23 and ST. 24 should be mandatory (the proposed selection of data headings is indicated in Annex 3; this selection is considered to be sufficient for identifying the sequence (listing) and for carrying out a good quality computerised search. **All the data elements which belong to the selected mandatory data headings are language-independent;**

(iii) **other data elements are optional** and may be useful for the evaluation

…/…

[2] This table - a copy of which is attached as Annex 2 - can be obtained from DNA Data Bank of Japan, Laboratory of Genetic Information Analysis, Center for Genetic Information Research, National Institute of Genetics, Mishina, Shizuoka
411 Japan; The European Molecular Biology Laboratory, Postfach 10.2209, D-69117 Heidelberg, Germany; NCBI/GenBank,
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

of the result of the computerised search and for the creation of a database entry for the Trilateral patent sequence database; most of these other data elements are also included in the DDBJ/EMBL/GenBank Feature Table.

8. Annex 4 contains a specimen SL for an application drawn up in French comprising:

(i) only the proposed mandatory elements

(ii) mandatory elements **and optional language-independent** (bibliographic) elements

(iii) mandatory elements, optional language-independent (bibliographic) elements, and optional language-independent (feature) elements **described** in the DDBJ/EMBL/GenBank Feature Table

(iv) mandatory elements, optional language-independent (bibliographic) elements, optional language-independent (feature) elements described in the DDBJ/EMBL/GenBank Feature Table and language-dependent elements **not** included in the Feature Table

(v) mandatory elements, optional language-independent (bibliographic) elements, optional language-independent (feature) elements described in the DDBJ/EMBL/GenBank Feature Table, language-dependent elements not included in the Feature Table, **and free text**.

From a comparison of the data under (iii) with alternatives (iv) and (v) in the above-mentioned Annex, it is clear that language-dependent text accounts for a only a tiny proportion of the information in the SL.

9. The language used in the search databases, e.g. the Trilateral Patent Sequence Database is English. This means that if any language-dependent text is supplied in a language other than English, it must be translated before being captured in the database. This is currently carried out by the sequence database producers and puts the Offices to extra expense.

…/…

Situation of the applicant

10. The existing situation is unsatisfactory for applicants, because if they wish to include language-dependent elements in the SL of a given application, they may have to make repeated alterations to the SLs originally encoded on computer in order to satisfy the language requirements of the various national/regional patent offices for second filings or for entry into the national/regional phase under the PCT1 and produce a corresponding number of diskettes.

Because of their technical complexity, drawing up SLs in different languages is time-consuming, and there is always a risk of errors occurring which may be prejudicial to the rights of the applicant.

11. Language-dependent technical terms are very similar from one language to another. For example, compare “ADN genomique” with “Genomic DNA” in Annex 4/iv or “Site de restriction ECURI” with “ECORI restriction site” in Annex 4/v. Demanding that such terms be translated is thus - arguably - over-formalistic, especially as the databases used in this field are generally exclusively in English.

In any case, it is expected that the DDBJ/EMBL/GenBank Feature Table will be updated from time to time to include new terms, leaving the number of language-dependent terms at a low rate.

III. PROPOSAL

12. The solution to the language problem addresses separately:

(a) the requirements for sequence listings submitted by the applicant during the international phase (Rules 5.2 and 13*ter*.1 PCT); and

(b) the requirements of the designated/elected Offices under Rule 13*ter*.2 PCT.

(a) International phase

13. It is proposed that on an optional basis for the applicant the language-dependent elements of sequence listings should be exempt from the principle that the entire application must be drafted in one and the same language, and

…/…

that they be accepted in English even if the rest of the application is in another language subject to the following conditions:

(i) the language-dependent elements must be kept to a minimum by using feature data from the DDBJ/EMBL/GenBank Feature Table and limiting the length of any free text (possibly 50 characters);

(ii) the definitions of the feature data included in the DDBJ/EMBL/GenBank Feature Table are available in the PCT languages.

(b) National/regional phase

14. Any designated/elected Office might require that the SL on entry into the national/regional phase be complemented with a glossary containing the translation (into the prescribed language) of the English language-dependent elements used in the SL.

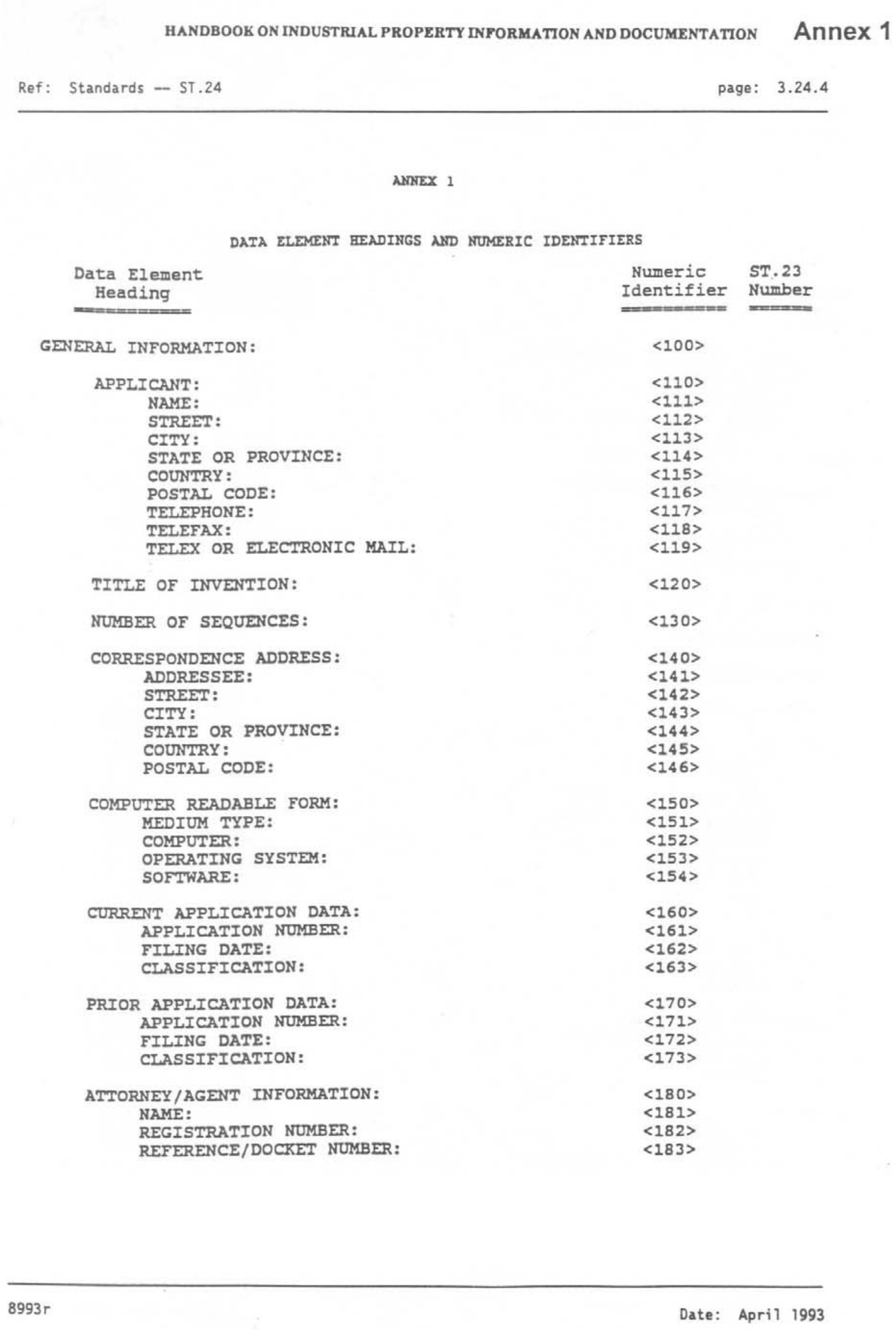
Annex 5 contains a specimen glossary produced for the example in Annex 4(v).

IV. CONCLUSIONS

15. The proposal brings the following advantages:

(a) To the extent that the applicant files a SL drawn up in English where the application is drawn up in another language, the ISAs get a SL in the language of the database and may thus proceed directly with the international search and capturing the SL in the database does not need any further translation.

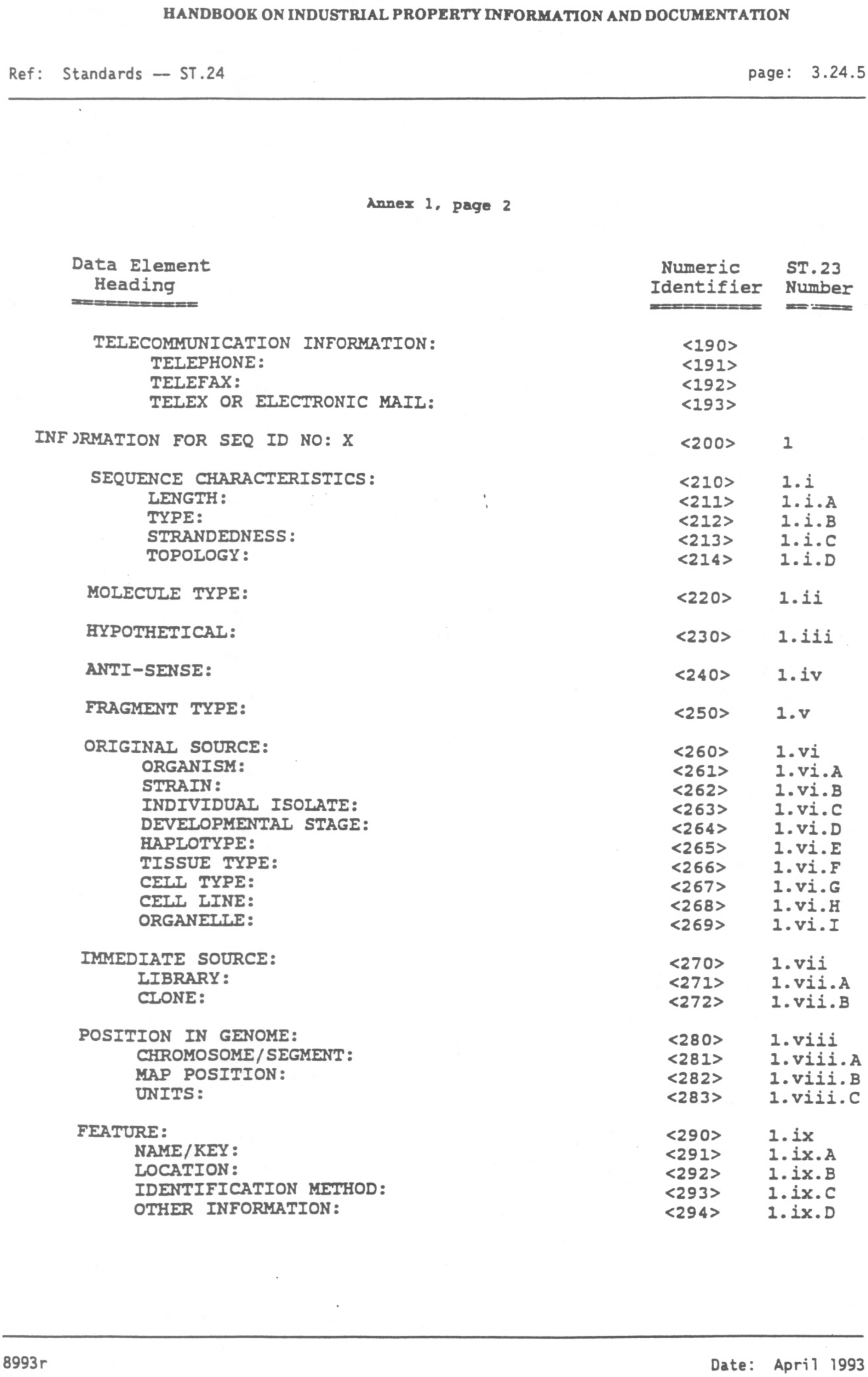
(b) Once the SL has been drawn up on paper and on diskette, the applicant can use it for any designated/elected Office provided that the SL on paper is supplemented by a glossary, if the designated/elected Office so requires.



ANNEX 1

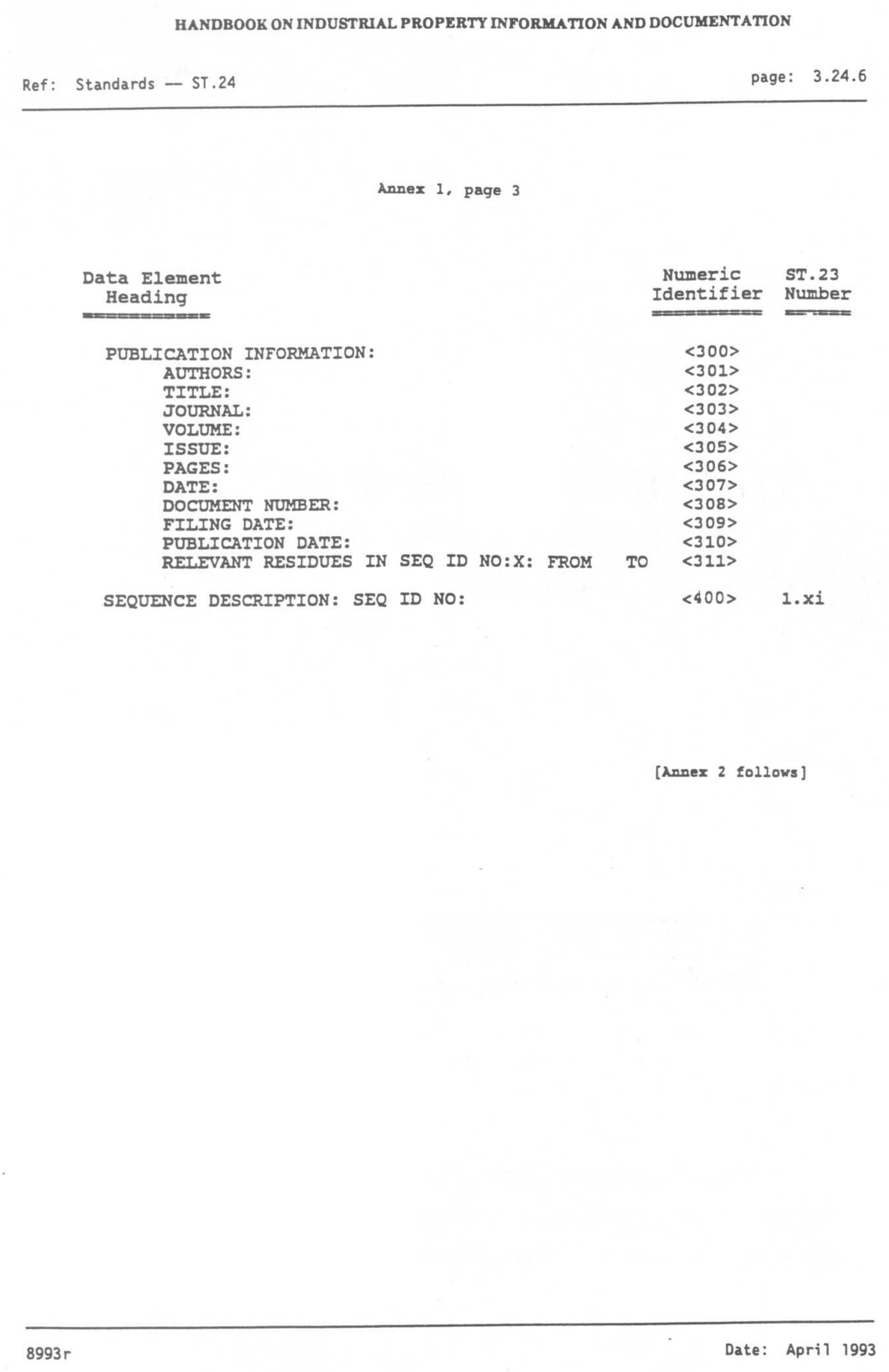
DATA ELEMENT HEADINGS AND NUMERIC IDENTIFIERS

Data Element Heading	Numeric Identifier	ST.23 Number
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>	



Annex 1, page 2

Data Element Heading	Numeric Identifier	ST.23 Number
TELECOMMUNICATION INFORMATION:	<190>	
TELEPHONE:	<191>	
TELEFAX:	<192>	
TELEX OR ELECTRONIC MAIL:	<193>	
INFORMATION FOR SEQ ID NO: X	<200>	1
SEQUENCE CHARACTERISTICS:	<210>	1.i
LENGTH:	<211>	1.i.A
TYPE:	<212>	1.i.B
STRANDEDNESS:	<213>	1.i.C
TOPOLOGY:	<214>	1.i.D
MOLECULE TYPE:	<220>	1.ii
HYPOTHETICAL:	<230>	1.iii
ANTI-SENSE:	<240>	1.iv
FRAGMENT TYPE:	<250>	1.v
ORIGINAL SOURCE:	<260>	1.vi
ORGANISM:	<261>	1.vi.A
STRAIN:	<262>	1.vi.B
INDIVIDUAL ISOLATE:	<263>	1.vi.C
DEVELOPMENTAL STAGE:	<264>	1.vi.D
HAPLOTYPE:	<265>	1.vi.E
TISSUE TYPE:	<266>	1.vi.F
CELL TYPE:	<267>	1.vi.G
CELL LINE:	<268>	1.vi.H
ORGANELLE:	<269>	1.vi.I
IMMEDIATE SOURCE:	<270>	1.vii
LIBRARY:	<271>	1.vii.A
CLONE:	<272>	1.vii.B
POSITION IN GENOME:	<280>	1.viii
CHROMOSOME/SEGMENT:	<281>	1.viii.A
MAP POSITION:	<282>	1.viii.B
UNITS:	<283>	1.viii.C
FEATURE:	<290>	1.ix
NAME/KEY:	<291>	1.ix.A
LOCATION:	<292>	1.ix.B
IDENTIFICATION METHOD:	<293>	1.ix.C
OTHER INFORMATION:	<294>	1.ix.D



Annex 1, page 3

Data Element Heading	Numeric Identifier	ST.23 Number
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>	1.xi

[Annex 2 follows]

Annex 2

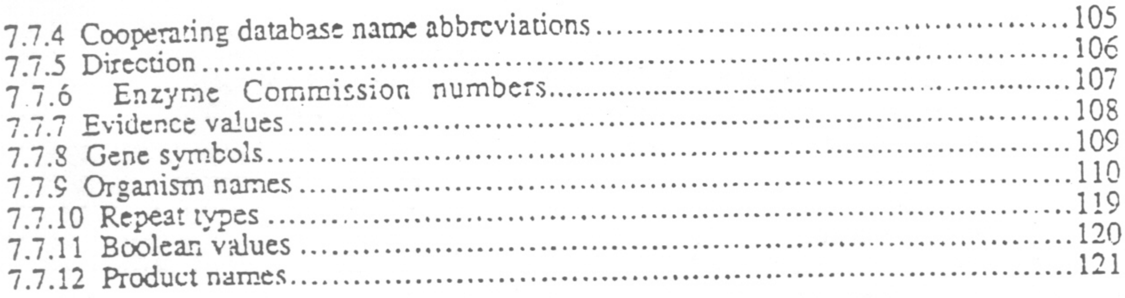
The DDBJ/EMBL/GenBank® Feature Table: Definition

Version 1.02
March 23, 1990

This version incorporates minor corrections to the GenBank new format feature table examples in Section 5 and Appendix L

DNA Data Bank of Japan, Mishima, Japan
EMBL Data Library, Heidelberg, Federal Republic of Germany
GenBank, Los Alamos, NM and Mountain View, CA, USA

Contents



The DDBJ/EMBL/GenBank® Feature Table:Definition

Version 1.02
March 23, 1990

> Das Unbeschreibliche,
> Hier ist's getan;
>
> Goethe
> *Faust, Part II*

1 Introduction

Nucleic acid sequences provide the fundamental starting point for describing and understanding the structure, function, and development of genetically diverse organisms. The primary structure of RNAs and proteins can be derived directly from the DNA or RNA genomic sequences. However, the translation, expression dependencies, and temporal sequence of events are neither straightforward nor well-defined. The GenBank, EMBL, and DDBJ nucleic acid sequence data banks have from their inception used tables of sites and features to describe the roles and locations of higher order sequence domains and elements within the genome of an organism. Appendix I shows examples of these feature tables. The independent development of feature table formats and annotation standards at EMBL and GenBank (later adopted by DDBJ) created significant difficulties for the data banks' data sharing efforts. As a result, in February, 1986, GenBank and EMBL began a collaborative effort (joined by DDBJ in 1987) to devise a common feature table format and common standards for annotation practice.

Early in this collaborative process it was recognized that both existing representational schemes suffered from similar inadequacies:

- Much of the information contained in the tables was difficult or impossible to extract by automatic methods.
- Sequences or features now recognized as important could not be annotated, but the tables' syntaxes limited their extension to include new classes of features.
- Features were not citable. They had no unique identifiers; hence, there could be no mechanism for linking between databases.
- The tables' syntaxes severely limited mechanisms for parsibly expressing complex features such as alternate splicing, circular molecules, read-through stop codons, sequence variation, multiple reading frames, secondary nucleic acid structures, pseudogenes, and other complex relationships between sequence regions.

2 Overview of the new Feature Table format

The overall goal of the new feature table design is to provide a more extensive vocabulary for describing features in a flexible framework for manipulating them. The range of features to be represented is diverse, including regions which:

(a) perform a biological function,
(b) affect or are the result of the expression of a biological function,
(c) interact with other molecules,
(d) affect replication of a sequence,
(e) affect or are the result of recombination of different sequences,
(f) are a recognizable repeated unit,
(g) have secondary or tertiary structure,
(h) exhibit variation, or
(i) have been revised or corrected.

The new feature table expands the feature vocabulary and adds new data items to allow more flexibility and a clearer specification of complex features. The format design, which is based on a tabular approach consists of the following items:

1. Feature key	a keyword indicating functional group
2. Location	instructions for finding the feature
3. Qualifiers	auxiliary information about a feature

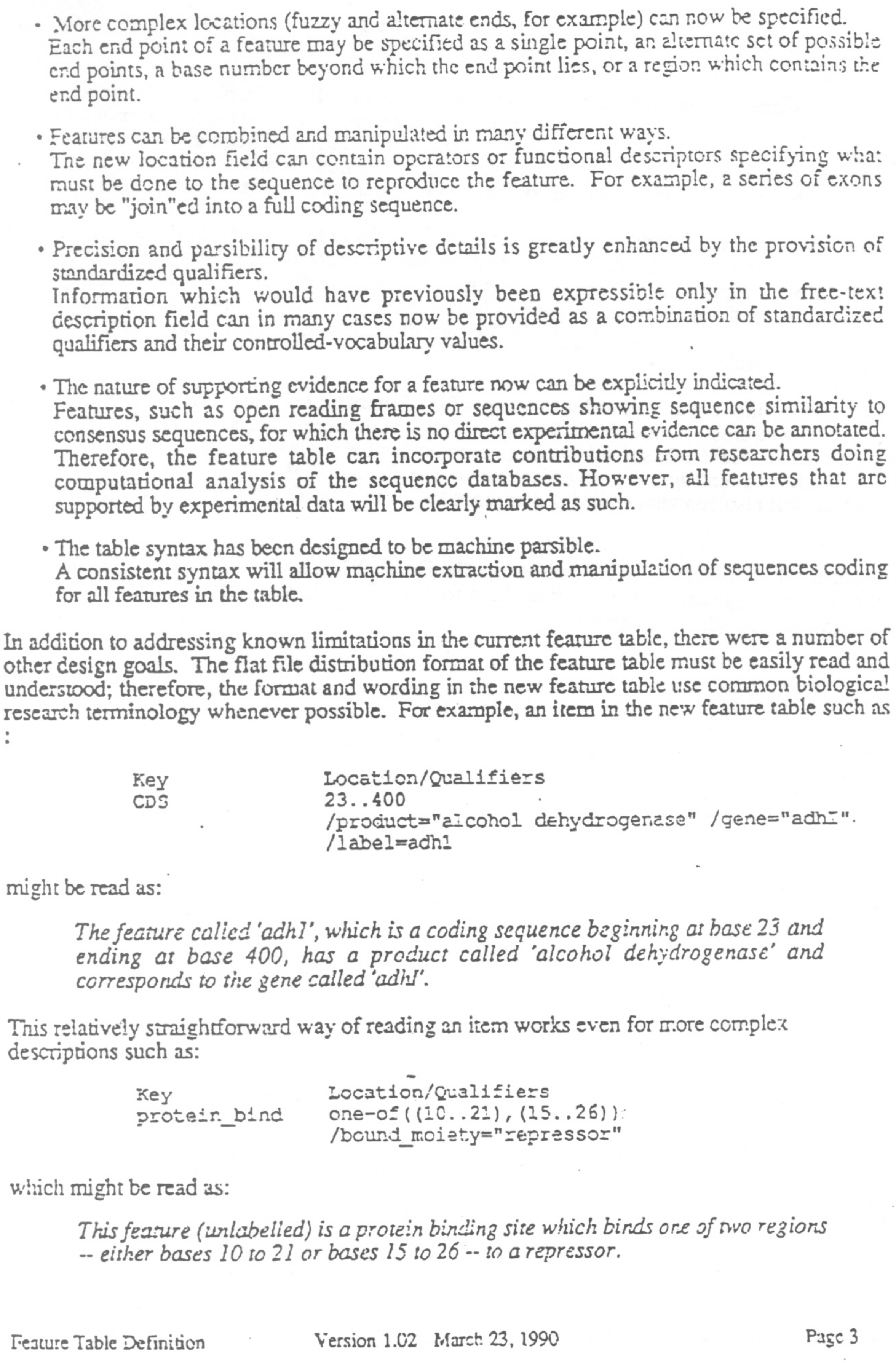
Each item will be discussed in more detail later; however, this design alleviates the limitations mentioned in the Introduction in the following ways:

- Features are distinct, citable entities.
  A new feature table data item -- the feature label (specified with a qualifier) -- permits direct identification of a feature. Therefore a feature can be referred to within an entry by feature label, between entries by primary accession number and feature label and between databases by database name, primary accession number and feature label. Note that this labelling convention allows cross-referencing to other databases (e.g., protein and genetic mapping databases as well as other nucleotide sequence databases) which adopt a compatible scheme of accession numbers and feature labels.

- A much richer set of keys allows specific annotation of important sequence features.
  Numerous features of biological significance (e.g., TATA and CAAT boxes; untranslated regions) which previously did not have distinct feature keys can now be clearly indicated.

- Related features can be easily specified and retrieved.
  Feature keys are arranged hierarchically, allowing complex and compound features to be expressed. Both location operators and the feature keys show feature relationships even when the features are not contiguous. Separate features, such as multiple mutations in a single strain, can be tied together using the "group" location operator. The hierarchy of feature keys allows broad categories of biological functionality, such as rRNAs, to be easily retrieved.

- Generic feature keys provide a means for entering new or undefined features.
  A number of "generic" or miscellaneous feature keys have been added to permit annotation of features that cannot be adequately described by existing feature keys. These generic feature keys will serve as an intermediate step in the identification and addition of new feature keys. The syntax has been designed to allow the addition of new feature keys as they are required.



- More complex locations (fuzzy and alternate ends, for example) can now be specified. Each end point of a feature may be specified as a single point, an alternate set of possible end points, a base number beyond which the end point lies, or a region which contains the end point.

- Features can be combined and manipulated in many different ways. The new location field can contain operators or functional descriptors specifying what must be done to the sequence to reproduce the feature. For example, a series of exons may be "join"ed into a full coding sequence.

- Precision and parsibility of descriptive details is greatly enhanced by the provision of standardized qualifiers. Information which would have previously been expressible only in the free-text description field can in many cases now be provided as a combination of standardized qualifiers and their controlled-vocabulary values.

- The nature of supporting evidence for a feature now can be explicitly indicated. Features, such as open reading frames or sequences showing sequence similarity to consensus sequences, for which there is no direct experimental evidence can be annotated. Therefore, the feature table can incorporate contributions from researchers doing computational analysis of the sequence databases. However, all features that are supported by experimental data will be clearly marked as such.

- The table syntax has been designed to be machine parsible. A consistent syntax will allow machine extraction and manipulation of sequences coding for all features in the table.

In addition to addressing known limitations in the current feature table, there were a number of other design goals. The flat file distribution format of the feature table must be easily read and understood; therefore, the format and wording in the new feature table use common biological research terminology whenever possible. For example, an item in the new feature table such as :

```
Key             Location/Qualifiers
CDS             23..400
                /product="alcohol dehydrogenase" /gene="adhI".
                /label=adh1
```

might be read as:

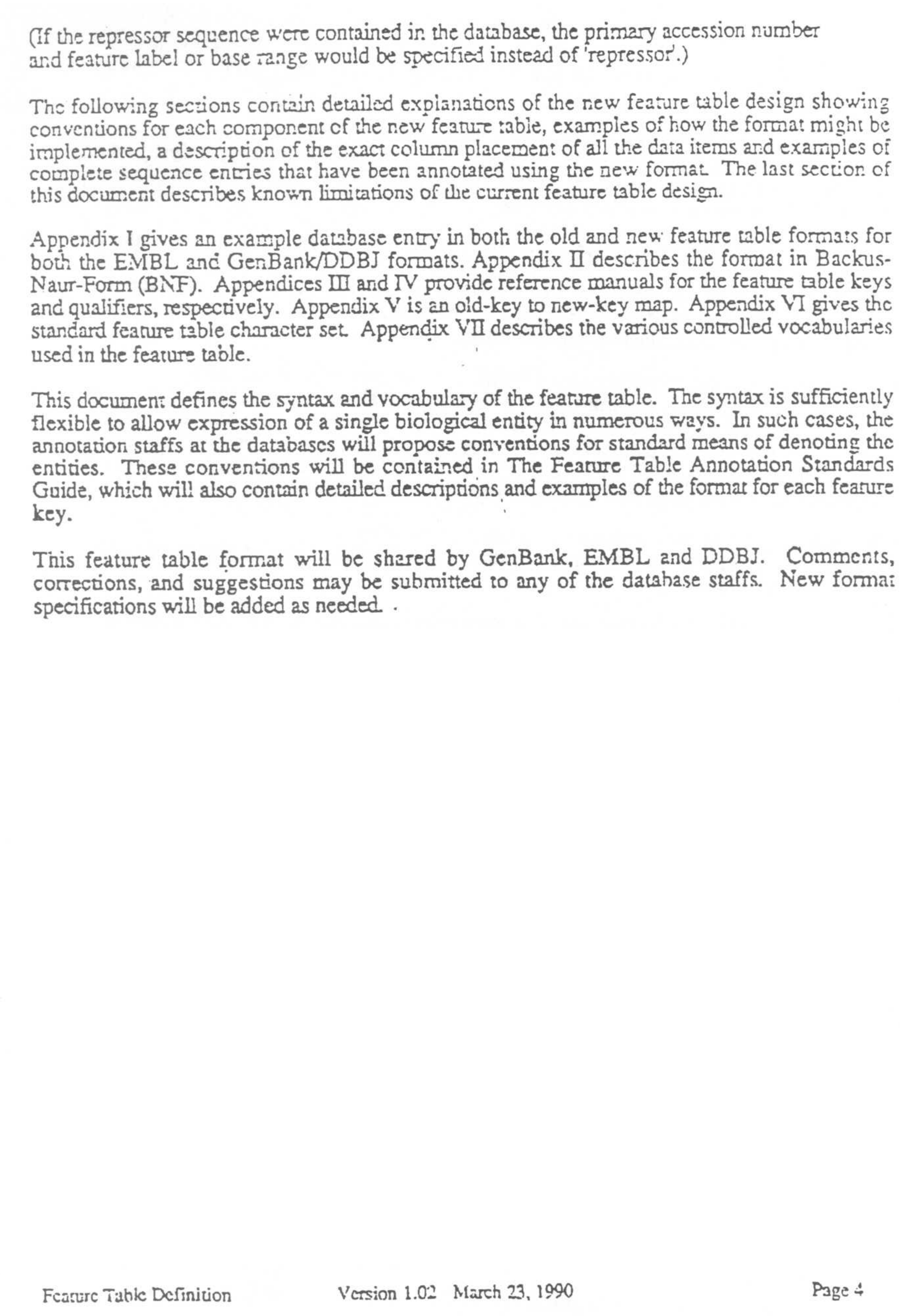
The feature called 'adh1', which is a coding sequence beginning at base 23 and ending at base 400, has a product called 'alcohol dehydrogenase' and corresponds to the gene called 'adhI'.

This relatively straightforward way of reading an item works even for more complex descriptions such as:

```
Key             Location/Qualifiers
protein_bind    one-of((10..21),(15..26))
                /bound_moiety="repressor"
```

which might be read as:

This feature (unlabelled) is a protein binding site which binds one of two regions -- either bases 10 to 21 or bases 15 to 26 -- to a repressor.



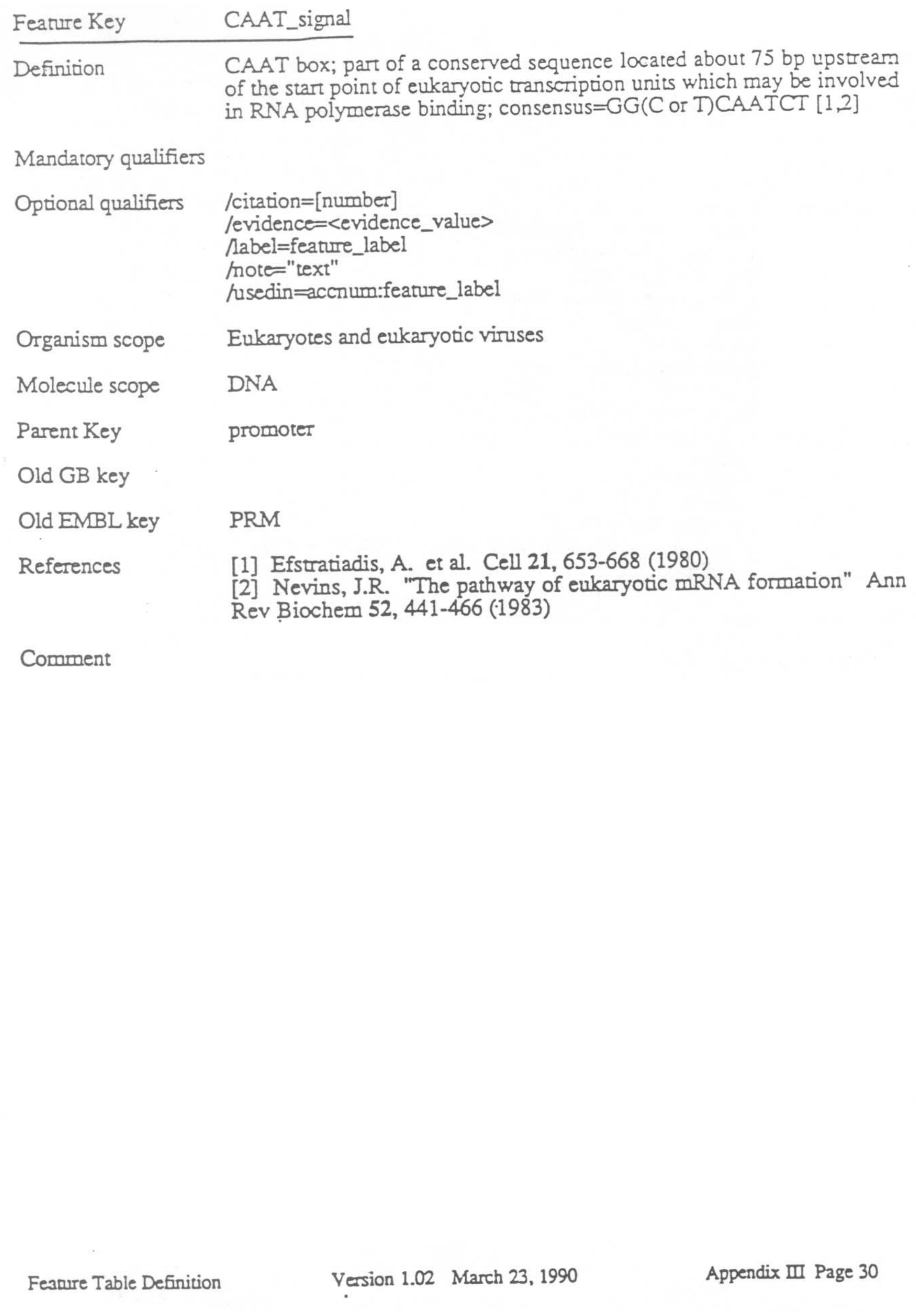
(If the repressor sequence were contained in the database, the primary accession number and feature label or base range would be specified instead of 'repressor'.)

The following sections contain detailed explanations of the new feature table design showing conventions for each component of the new feature table, examples of how the format might be implemented, a description of the exact column placement of all the data items and examples of complete sequence entries that have been annotated using the new format. The last section of this document describes known limitations of the current feature table design.

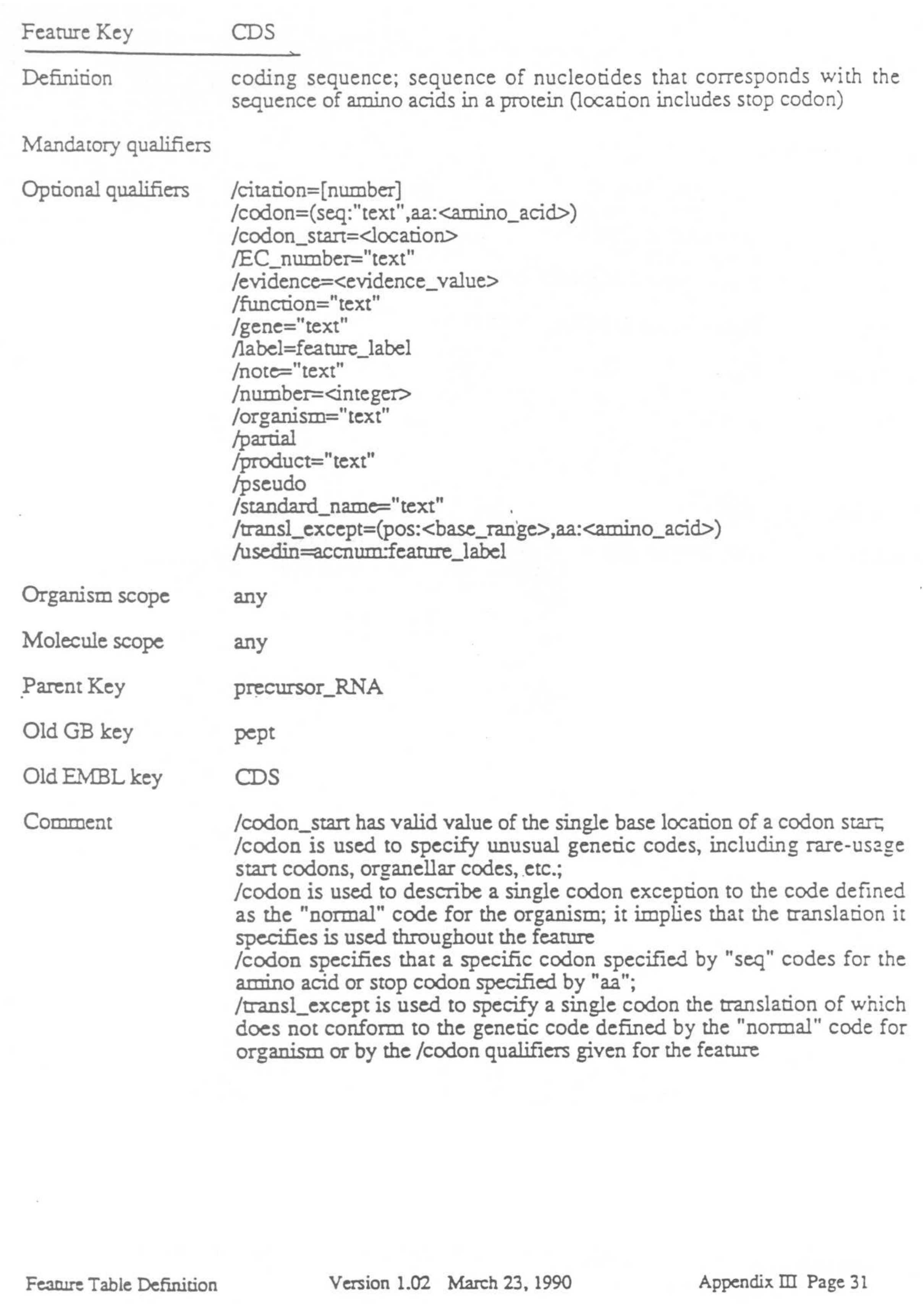
Appendix I gives an example database entry in both the old and new feature table formats for both the EMBL and GenBank/DDBJ formats. Appendix II describes the format in Backus-Naur-Form (BNF). Appendices III and IV provide reference manuals for the feature table keys and qualifiers, respectively. Appendix V is an old-key to new-key map. Appendix VI gives the standard feature table character set. Appendix VII describes the various controlled vocabularies used in the feature table.

This document defines the syntax and vocabulary of the feature table. The syntax is sufficiently flexible to allow expression of a single biological entity in numerous ways. In such cases, the annotation staffs at the databases will propose conventions for standard means of denoting the entities. These conventions will be contained in The Feature Table Annotation Standards Guide, which will also contain detailed descriptions and examples of the format for each feature key.

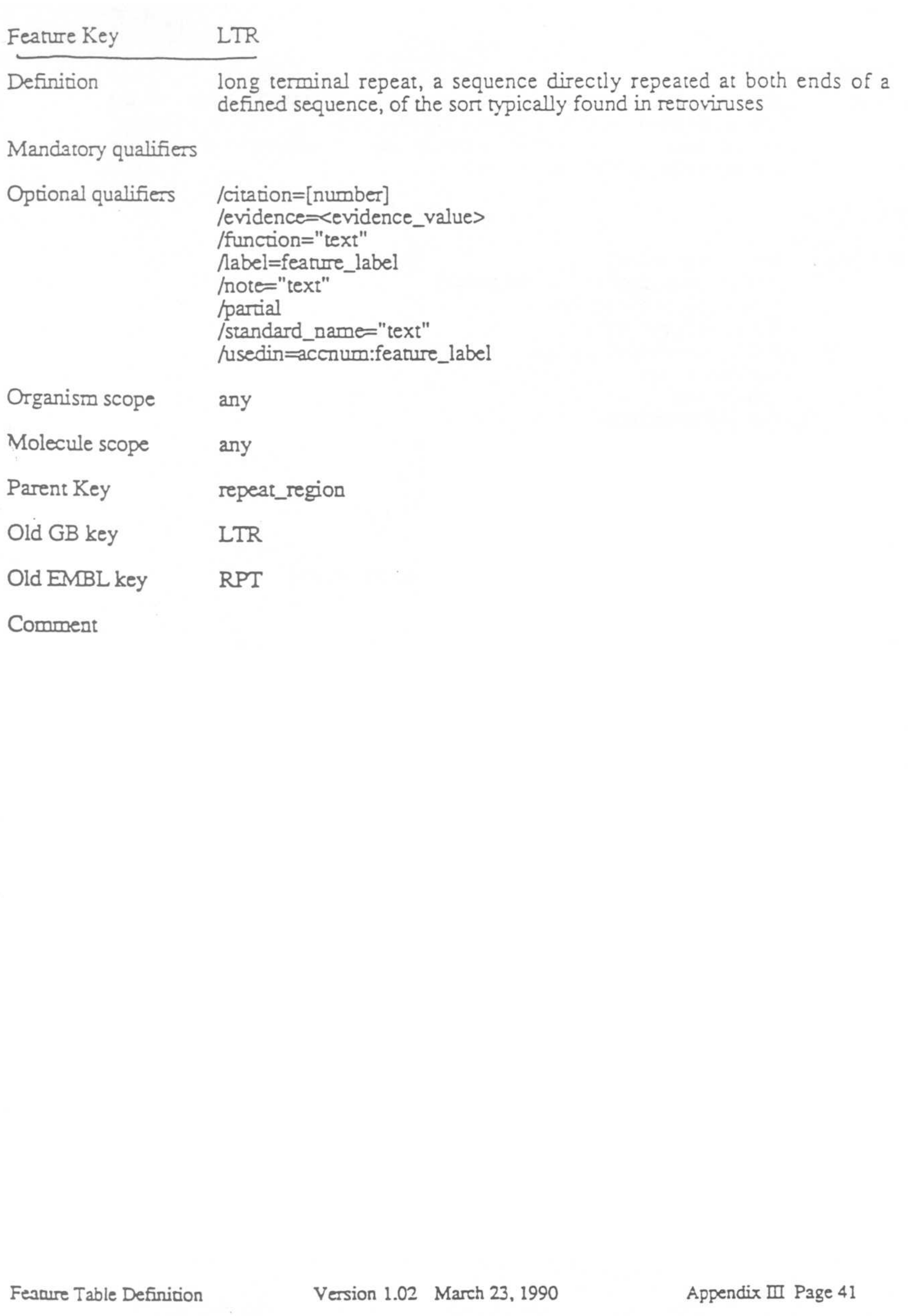
This feature table format will be shared by GenBank, EMBL and DDBJ. Comments, corrections, and suggestions may be submitted to any of the database staffs. New format specifications will be added as needed.



Feature Key	CAAT_signal
Definition	CAAT box; part of a conserved sequence located about 75 bp upstream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG(C or T)CAATCT [1,2]
Mandatory qualifiers	
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes and eukaryotic viruses
Molecule scope	DNA
Parent Key	promoter
Old GB key	
Old EMBL key	PRM
References	[1] Efstratiadis, A. et al. Cell **21**, 653-668 (1980) [2] Nevins, J.R. "The pathway of eukaryotic mRNA formation" Ann Rev Biochem **52**, 441-466 (1983)
Comment	

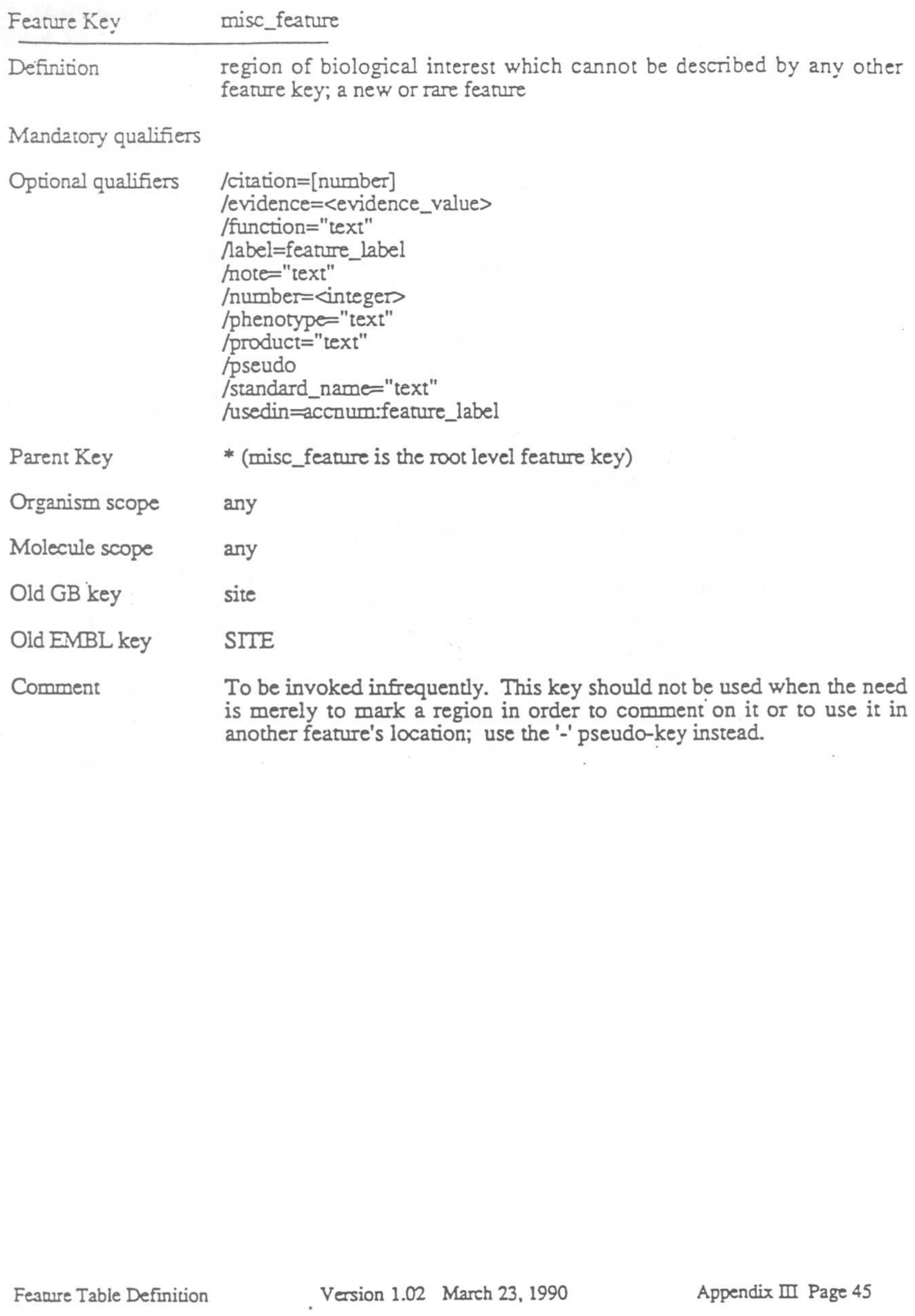


Feature Key	CDS
Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon)
Mandatory qualifiers	
Optional qualifiers	/citation=[number] /codon=(seq:"text",aa:<amino_acid>) /codon_start=<location> /EC_number="text" /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /number=<integer> /organism="text" /partial /product="text" /pseudo /standard_name="text" /transl_except=(pos:<base_range>,aa:<amino_acid>) /usedin=accnum:feature_label
Organism scope	any
Molecule scope	any
Parent Key	precursor_RNA
Old GB key	pept
Old EMBL key	CDS
Comment	/codon_start has valid value of the single base location of a codon start; /codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.; /codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature /codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa"; /transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature

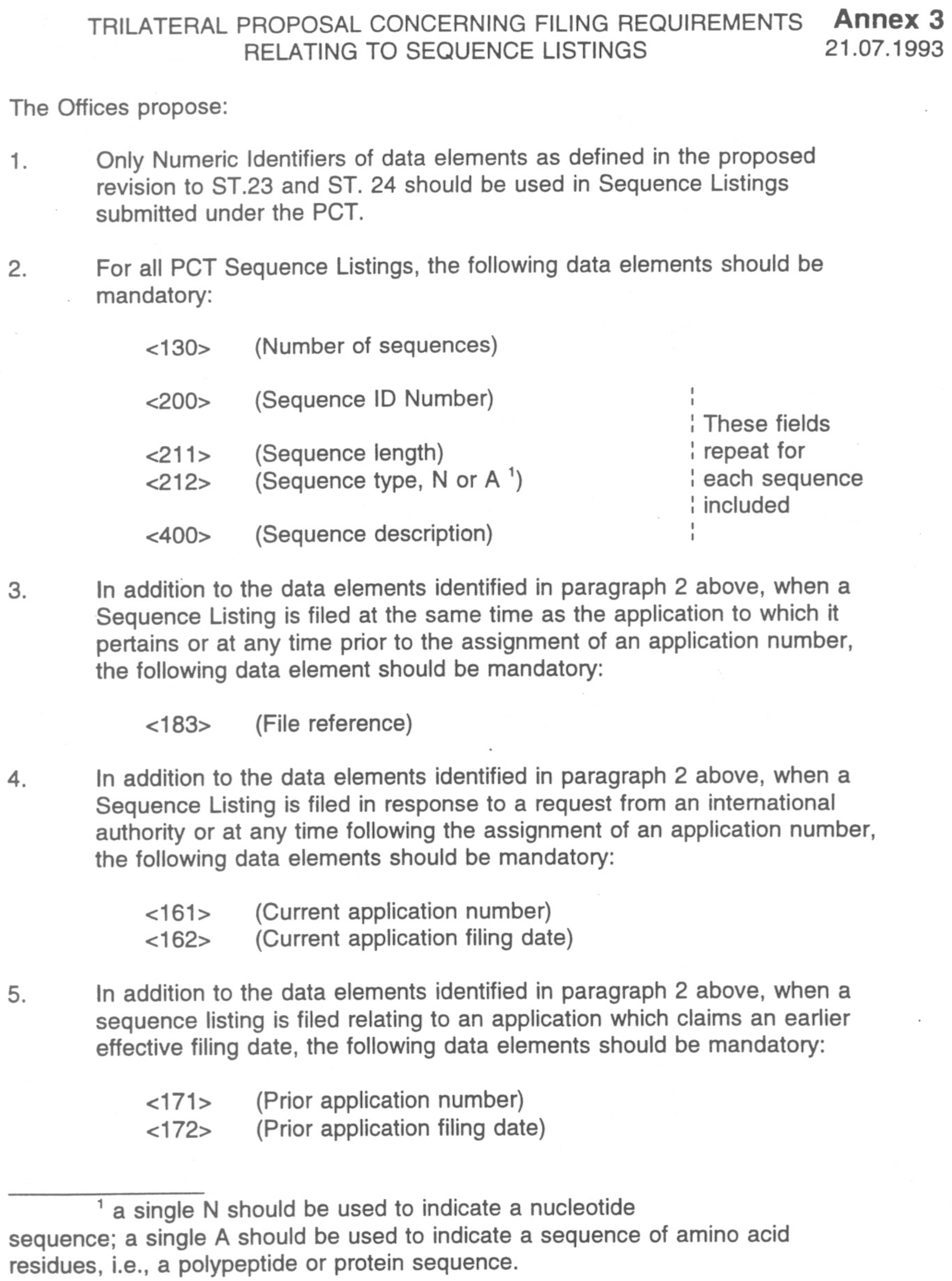


Feature Key	LTR
Definition	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
Mandatory qualifiers	
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /partial /standard_name="text" /usedin=accnum:feature_label
Organism scope	any
Molecule scope	any
Parent Key	repeat_region
Old GB key	LTR
Old EMBL key	RPT
Comment	

Feature Key	mat_peptide
Definition	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)
Mandatory qualifiers	
Optional qualifiers	/citation=[number] /codon=(seq:"text",aa:<amino_acid>) /codon_start=<location> /EC_number="text" /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /organism="text" /partial /product="text" /pseudo /standard_name="text" /transl_except=(pos:<base_range>,aa:<amino_acid>) /usedin=accnum:feature_label
Organism scope	any
Molecule scope	any
Parent Key	CDS
Old GB key	matp; pept
Old EMBL key	CDS
Comment	/codon_start has valid value of the single base location of a codon start; /codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.; /codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature /codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa"; /transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature



Feature Key	misc_feature
Definition	region of biological interest which cannot be described by any other feature key; a new or rare feature
Mandatory qualifiers	
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /number=<integer> /phenotype="text" /product="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Parent Key	* (misc_feature is the root level feature key)
Organism scope	any
Molecule scope	any
Old GB key	site
Old EMBL key	SITE
Comment	To be invoked infrequently. This key should not be used when the need is merely to mark a region in order to comment on it or to use it in another feature's location; use the '-' pseudo-key instead.



TRILATERAL PROPOSAL CONCERNING FILING REQUIREMENTS RELATING TO SEQUENCE LISTINGS

Annex 3
21.07.1993

The Offices propose:

1. Only Numeric Identifiers of data elements as defined in the proposed revision to ST.23 and ST. 24 should be used in Sequence Listings submitted under the PCT.

2. For all PCT Sequence Listings, the following data elements should be mandatory:

<130>	(Number of sequences)	
<200>	(Sequence ID Number)	These fields repeat for each sequence included
<211>	(Sequence length)	
<212>	(Sequence type, N or A [1])	
<400>	(Sequence description)	

3. In addition to the data elements identified in paragraph 2 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 should be mandatory:

<183>	(File reference)

4. In addition to the data elements identified in paragraph 2 above, when a Sequence Listing is filed in response to a request from an international authority or at any time following the assignment of an application number, the following data elements should be mandatory:

<161>	(Current application number)
<162>	(Current application filing date)

5. In addition to the data elements identified in paragraph 2 above, when a sequence listing is filed relating to an application which claims an earlier effective filing date, the following data elements should be mandatory:

<171>	(Prior application number)
<172>	(Prior application filing date)

[1] a single N should be used to indicate a nucleotide sequence; a single A should be used to indicate a sequence of amino acid residues, i.e., a polypeptide or protein sequence.

6. All data elements mentioned above (mandatory elements) should contain only unaccented upper case characters "A" through "Z", unaccented lower case characters "a" through "z", the numerals "0" through "9" and the special characters " ", "!", """, "#", "$", "%", "&", "'", "(", ")", "*", "+", ",", "-", ".", "/", ":", ";", "<", "=", ">", "?", "@", "[", "\", "]", "^", "_", "`", "{", "|", "}", "~" and "⌂". This set of characters can most easily be defined as those characters occupying columns 2 through 7 of the compatible code pages referred to in paragraph 6 of ST. 24. This is the same set of characters defined in columns 2 through 7 of ISO 646.

7. All data elements defined in the proposed revision to ST.23 and ST.24, not mentioned above, should be optional. When such optional data elements are presented, they should be presented in accordance with paragraph 6 of ST.24.

8. As far as the language element is concerned, it is desirable to use only English.

9. Sequence Listings presented in computer-readable form should be presented on standard density (720Kb), three and one-half inch flexible disks using PC-DOS or MS-DOS operating systems only.

10. The Offices consider the above identified non-sequence data elements to be necessary or advisable for the following reasons:

<130> The number of sequences is used as a check to determine whether all sequences that were intended to be included on the Sequence Listing or its computer-readable form are, in fact, present.

<161> The application number is used not only to associate the Sequence Listing with its application, but also in search systems to identify the results of a search.

<162> The application filing date is used in search systems and by examiners to determine the date which must be anticipated.

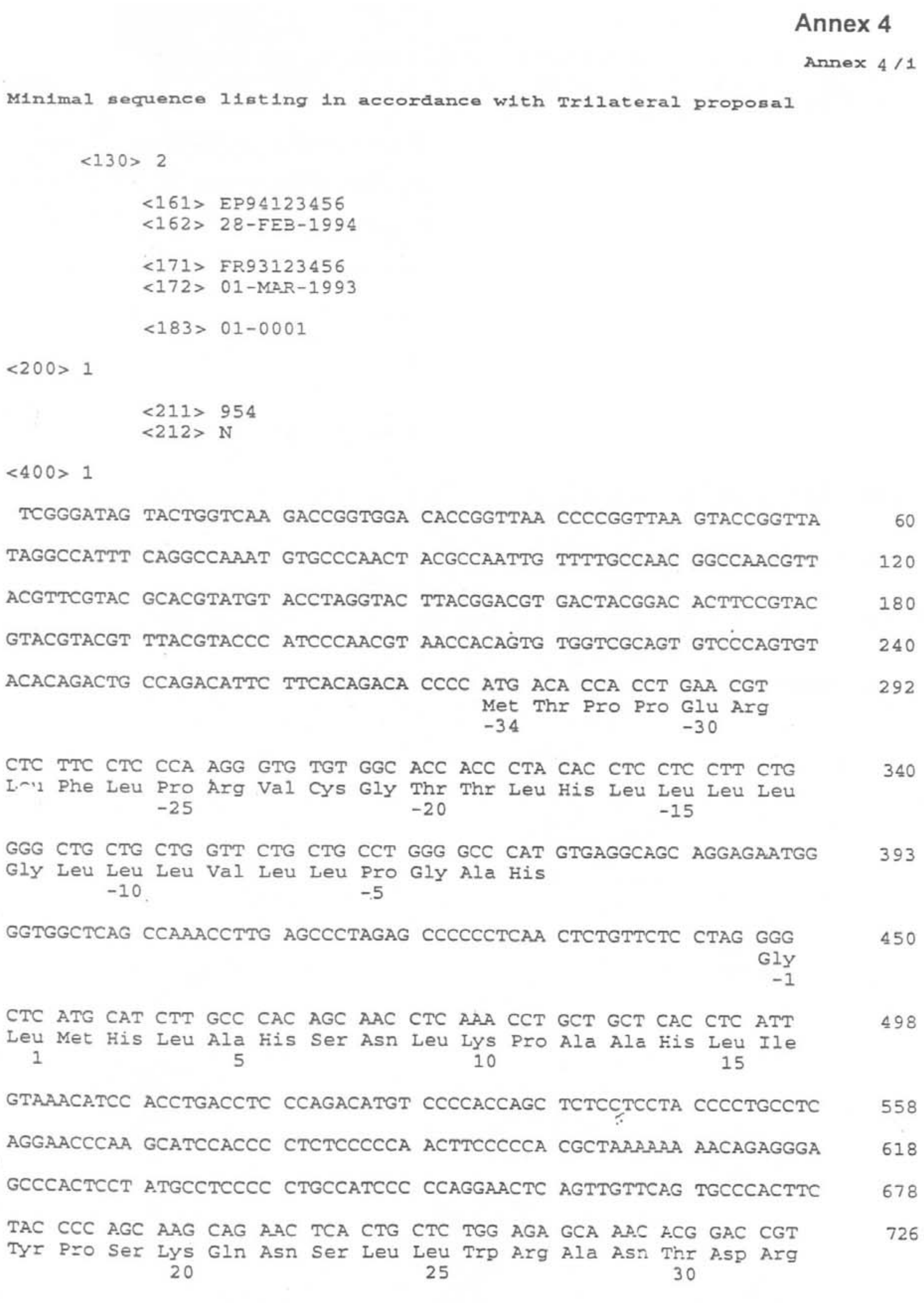
<171> The prior application number is used in search systems to determine whether a search has already been performed on the prior application.

<172> The prior application filing date is used in search systems and by examiners to determine the date which must be anticipated.

<183> the file reference is the same as that discussed in Section 109 of the Administrative Instructions. It is used to associate the Sequence Listing or its computer-readable form with the application.

<211> The sequence length is used as a check to determine whether all elements of an intended sequence are, in fact, present.

<212> The sequence type is used to determine whether the sequence has been properly presented, that is, whether single letter codes have been inadvertently used for amino acid residues since the overlap with the IUPAC code for nucleic acids is substantial. The entry for this data element can be limited to "N" for nucleotide or "A" for amino acid residues thus making this element language neutral.



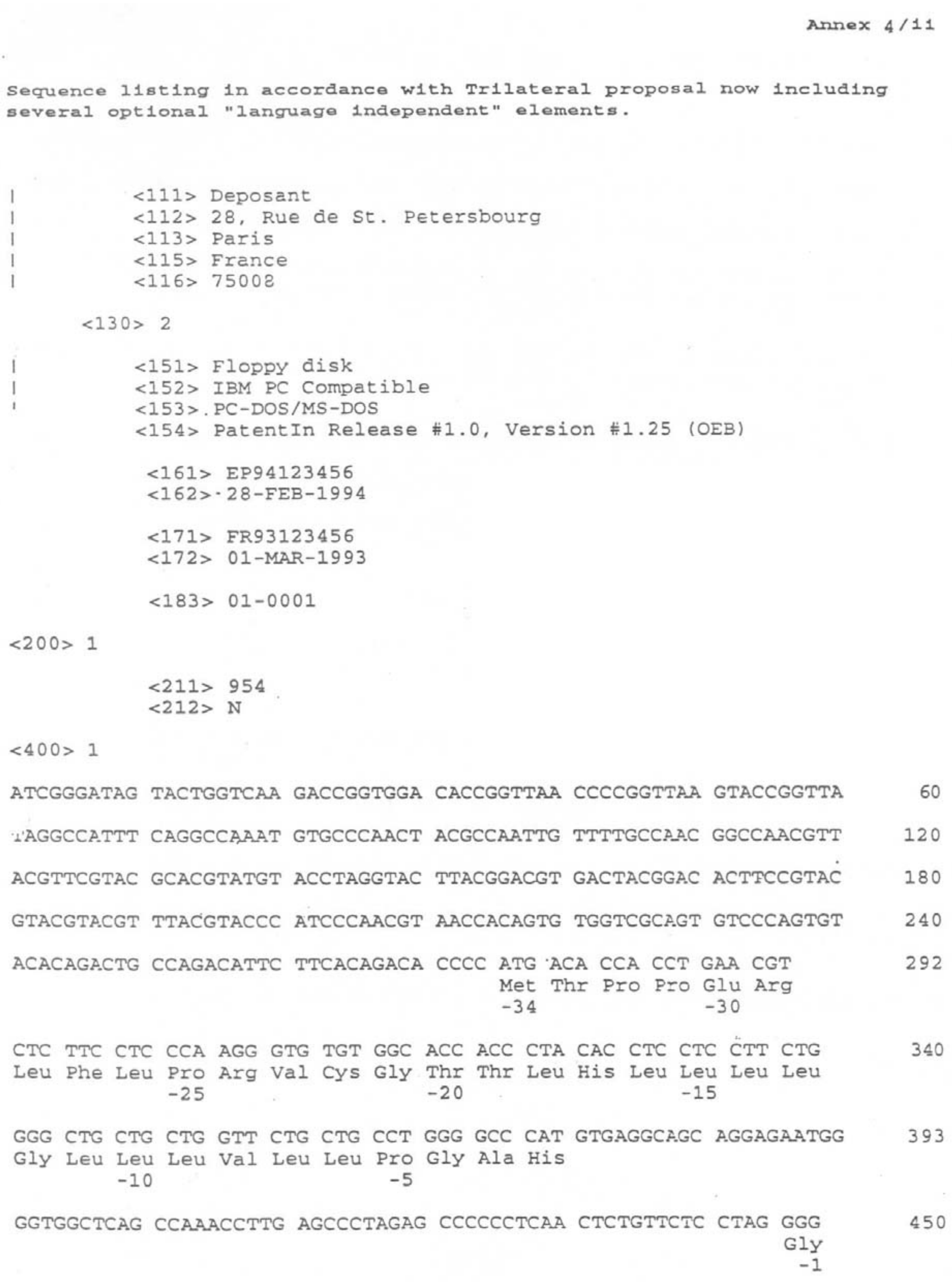
Annex 4

Annex 4 /1

Minimal sequence listing in accordance with Trilateral proposal

```
     <130> 2

          <161> EP94123456
          <162> 28-FEB-1994

          <171> FR93123456
          <172> 01-MAR-1993

          <183> 01-0001

<200> 1

          <211> 954
          <212> N

<400> 1

 TCGGGATAG TACTGGTCAA GACCGGTGGA CACCGGTTAA CCCCGGTTAA GTACCGGTTA        60

TAGGCCATTT CAGGCCAAAT GTGCCCAACT ACGCCAATTG TTTTGCCAAC GGCCAACGTT       120

ACGTTCGTAC GCACGTATGT ACCTAGGTAC TTACGGACGT GACTACGGAC ACTTCCGTAC       180

GTACGTACGT TTACGTACCC ATCCCAACGT AACCACAGTG TGGTCGCAGT GTCCCAGTGT       240

ACACAGACTG CCAGACATTC TTCACAGACA CCCC ATG ACA CCA CCT GAA CGT           292
                                      Met Thr Pro Pro Glu Arg
                                      -34             -30

CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC CTT CTG         340
Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu Leu Leu
            -25                 -20                 -15

GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAT GTGAGGCAGC AGGAGAATGG       393
Gly Leu Leu Leu Val Leu Leu Pro Gly Ala His
        -10                 -5

GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA CTCTGTTCTC CTAG GGG         450
                                                            Gly
                                                             -1

CTC ATG CAT CTT GCC CAC AGC AAC CTC AAA CCT GCT GCT CAC CTC ATT         498
Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His Leu Ile
  1               5                  10                  15

GTAAACATCC ACCTGACCTC CCAGACATGT CCCCACCAGC TCTCCTCCTA CCCCTGCCTC       558

AGGAACCCAA GCATCCACCC CTCTCCCCCA ACTTCCCCCA CGCTAAAAAA AACAGAGGGA       618

GCCCACTCCT ATGCCTCCCC CTGCCATCCC CCAGGAACTC AGTTGTTCAG TGCCCACTTC       678

TAC CCC AGC AAG CAG AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT         726
Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg
            20                  25                  30
```

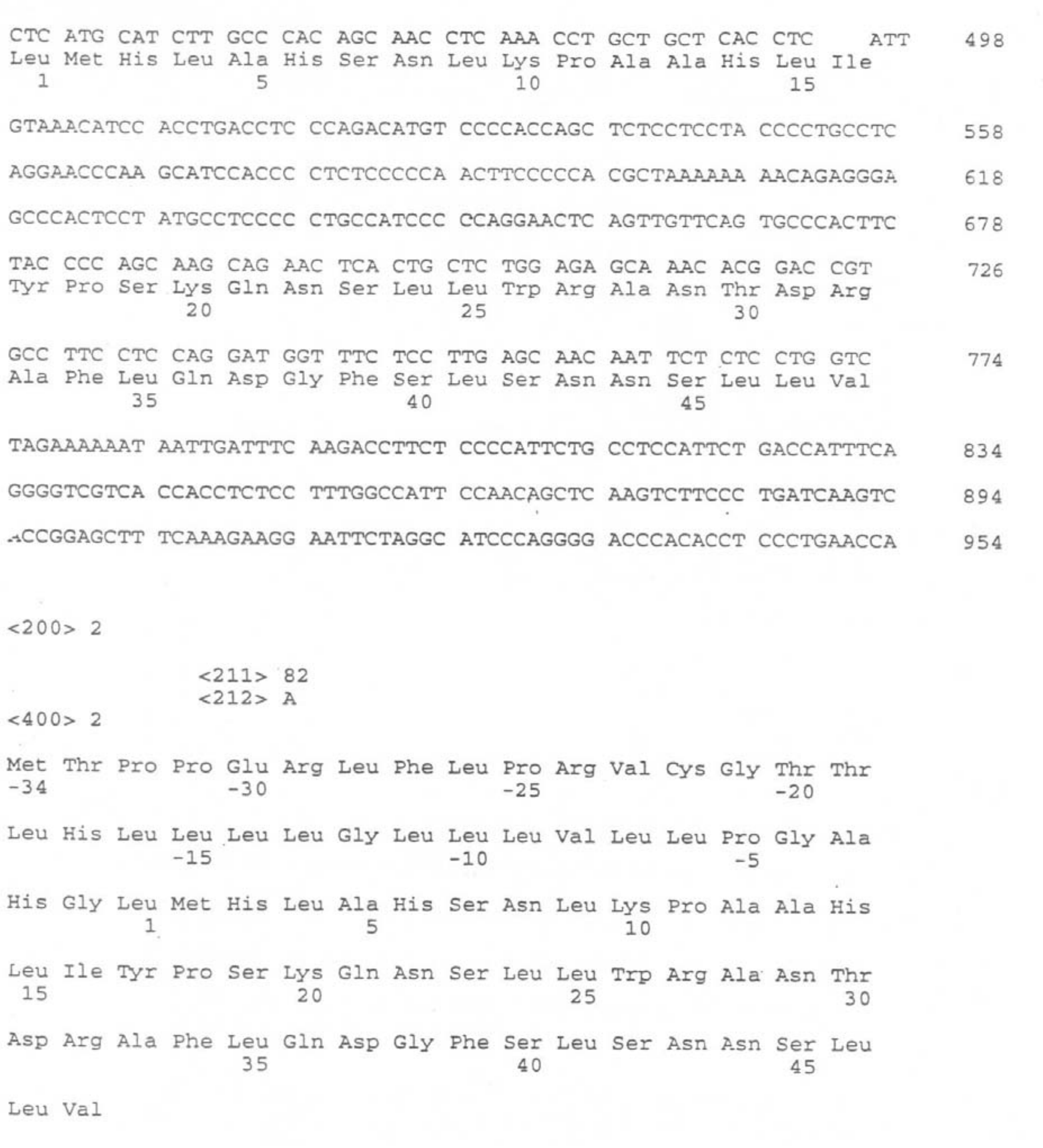
```
GCC TTC CTC CAG GAT GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC      774
Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val
         35                  40                  45

TAGAAAAAAT AATTGATTTC AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA    834

GGGGTCGTCA CCACCTCTCC TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC    894

ACCGGAGCTT TCAAAGAAGG AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA    954


<200> 2

              <211> 82
              <212> A
<400> 2

Met Thr Pro Pro Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr
-34             -30                 -25                 -20

Leu His Leu Leu Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala
            -15                 -10                 -5

His Gly Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His
         1               5                  10

Leu Ile Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr
 15                  20                  25                  30

Asp Arg Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu
                 35                  40                  45

Leu Val
```



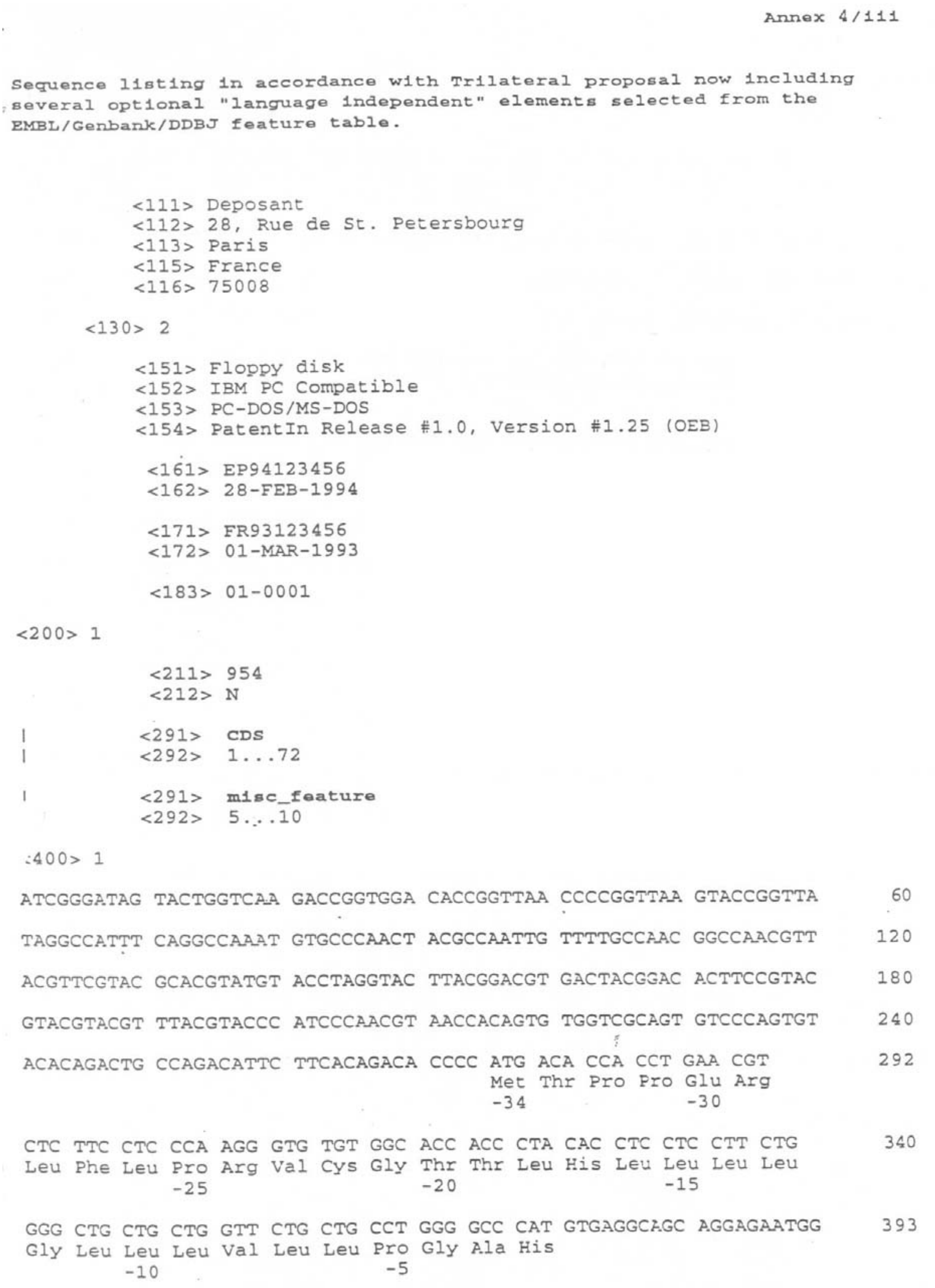
Annex 4/11

Sequence listing in accordance with Trilateral proposal now including several optional "language independent" elements.

```
|          <111> Deposant
|          <112> 28, Rue de St. Petersbourg
|          <113> Paris
|          <115> France
|          <116> 75008

      <130> 2

|          <151> Floppy disk
|          <152> IBM PC Compatible
'          <153> PC-DOS/MS-DOS
           <154> PatentIn Release #1.0, Version #1.25 (OEB)

            <161> EP94123456
            <162> 28-FEB-1994

            <171> FR93123456
            <172> 01-MAR-1993

            <183> 01-0001

<200> 1

            <211> 954
            <212> N

<400> 1

ATCGGGATAG TACTGGTCAA GACCGGTGGA CACCGGTTAA CCCCGGTTAA GTACCGGTTA      60

TAGGCCATTT CAGGCCAAAT GTGCCCAACT ACGCCAATTG TTTTGCCAAC GGCCAACGTT     120

ACGTTCGTAC GCACGTATGT ACCTAGGTAC TTACGGACGT GACTACGGAC ACTTCCGTAC     180

GTACGTACGT TTACGTACCC ATCCCAACGT AACCACAGTG TGGTCGCAGT GTCCCAGTGT     240

ACACAGACTG CCAGACATTC TTCACAGACA CCCC ATG ACA CCA CCT GAA CGT         292
                                      Met Thr Pro Pro Glu Arg
                                      -34             -30

CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC CTT CTG       340
Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu Leu Leu
            -25                 -20                 -15

GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAT GTGAGGCAGC AGGAGAATGG     393
Gly Leu Leu Leu Val Leu Leu Pro Gly Ala His
        -10                 -5

GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA CTCTGTTCTC CTAG GGG       450
                                                            Gly
                                                             -1
```



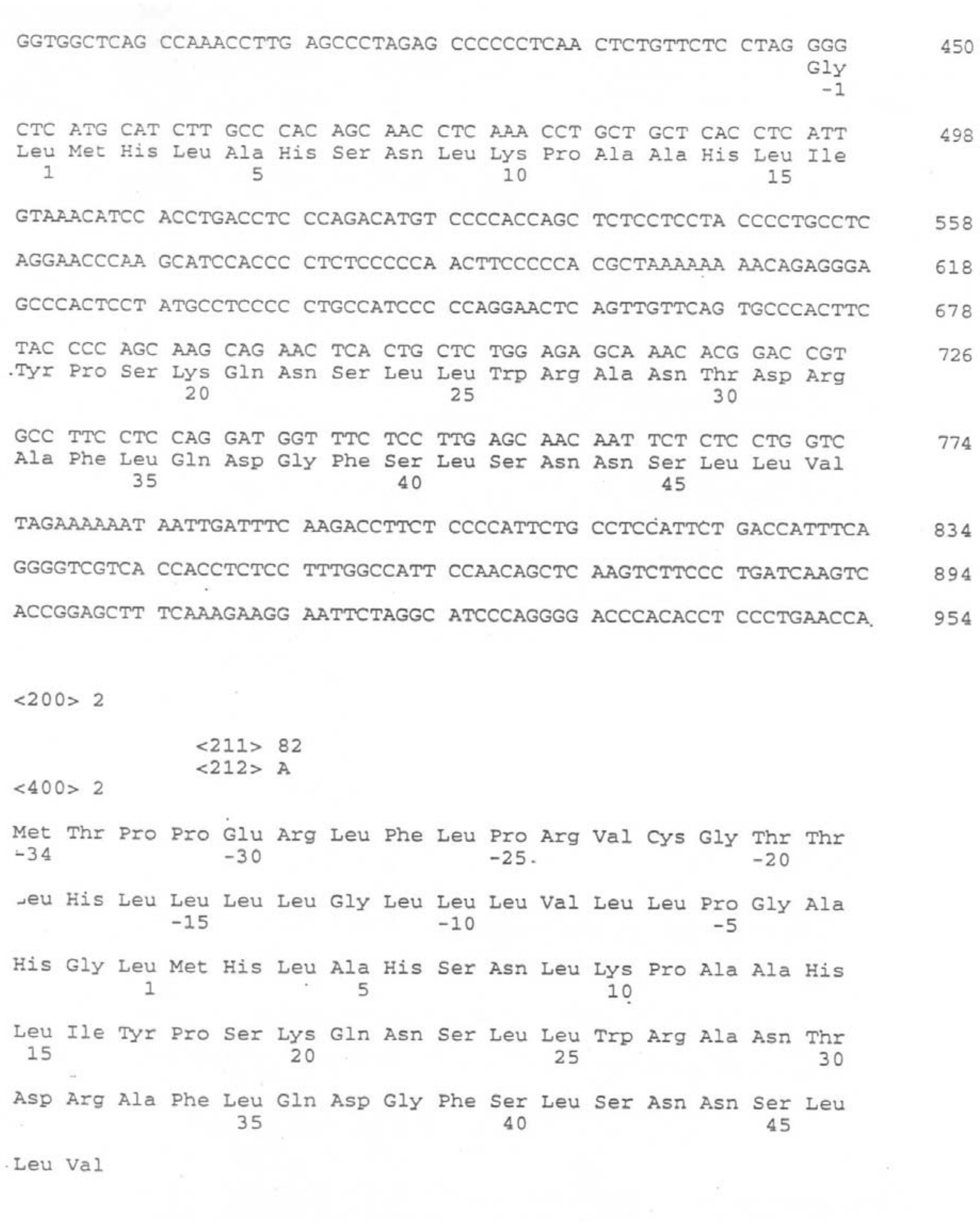
```
CTC ATG CAT CTT GCC CAC AGC AAC CTC AAA CCT GCT GCT CAC CTC    ATT     498
Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His Leu Ile
  1               5                   10                  15

GTAAACATCC ACCTGACCTC CCAGACATGT CCCCACCAGC TCTCCTCCTA CCCCTGCCTC      558

AGGAACCCAA GCATCCACCC CTCTCCCCCA ACTTCCCCCA CGCTAAAAAA AACAGAGGGA      618

GCCCACTCCT ATGCCTCCCC CTGCCATCCC CCAGGAACTC AGTTGTTCAG TGCCCACTTC      678

TAC CCC AGC AAG CAG AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT        726
Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg
             20                  25                  30

GCC TTC CTC CAG GAT GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC        774
Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val
         35                  40                  45

TAGAAAAAAT AATTGATTTC AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA      834

GGGGTCGTCA CCACCTCTCC TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC      894

ACCGGAGCTT TCAAAGAAGG AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA      954


<200> 2

              <211> 82
              <212> A
<400> 2

Met Thr Pro Pro Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr
-34             -30                 -25                 -20

Leu His Leu Leu Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala
            -15                 -10                  -5

His Gly Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His
             1               5                  10

Leu Ile Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr
 15                 20                  25                  30

Asp Arg Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu
                35                  40                  45

Leu Val
```



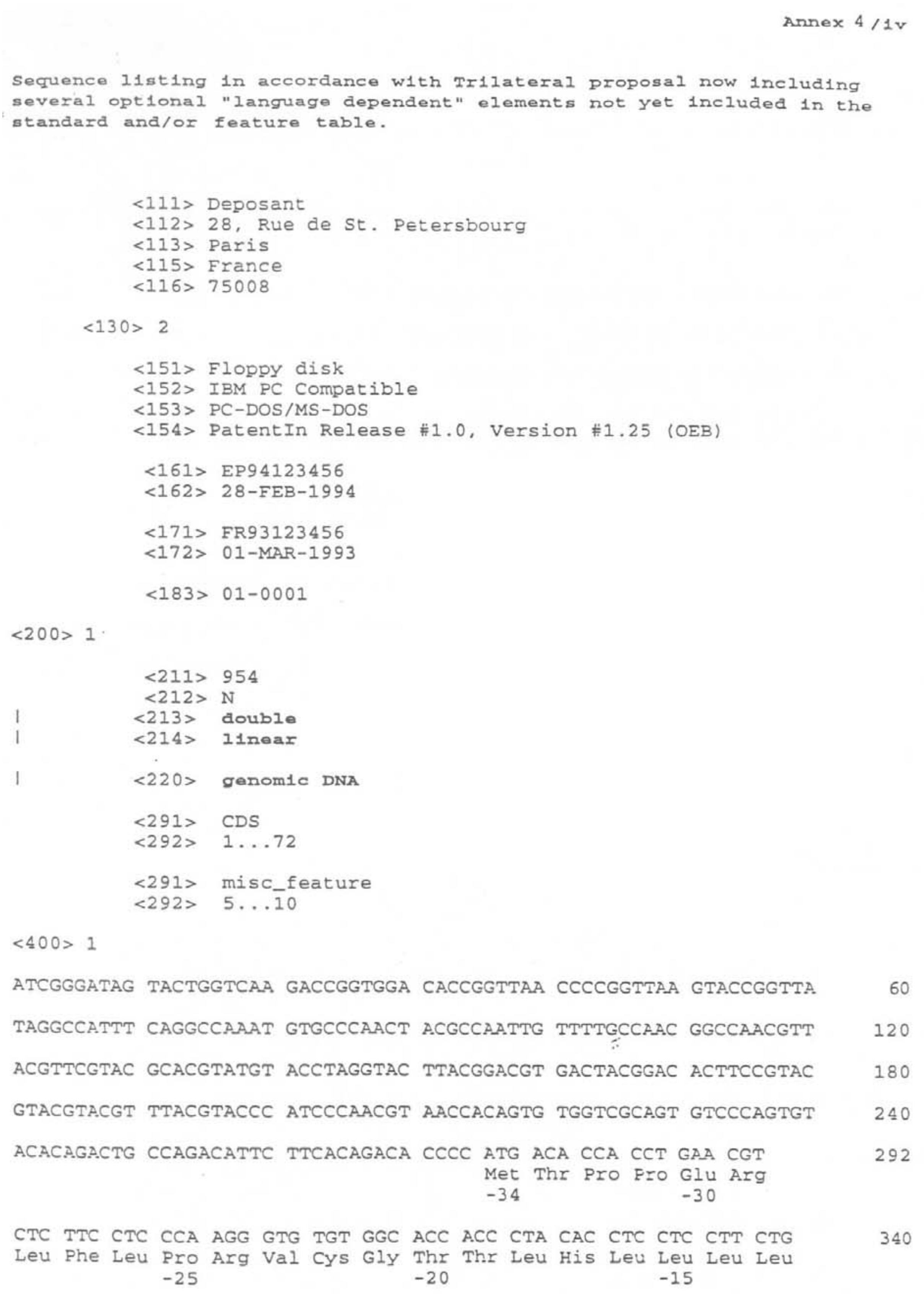
Annex 4/iii

Sequence listing in accordance with Trilateral proposal now including several optional "language independent" elements selected from the EMBL/Genbank/DDBJ feature table.

```
          <111> Deposant
          <112> 28, Rue de St. Petersbourg
          <113> Paris
          <115> France
          <116> 75008

      <130> 2

          <151> Floppy disk
          <152> IBM PC Compatible
          <153> PC-DOS/MS-DOS
          <154> PatentIn Release #1.0, Version #1.25 (OEB)

           <161> EP94123456
           <162> 28-FEB-1994

           <171> FR93123456
           <172> 01-MAR-1993

           <183> 01-0001

<200> 1

           <211> 954
           <212> N

|         <291>  CDS
|         <292>  1...72

|         <291>  misc_feature
          <292>  5...10

<400> 1

ATCGGGATAG TACTGGTCAA GACCGGTGGA CACCGGTTAA CCCCGGTTAA GTACCGGTTA      60

TAGGCCATTT CAGGCCAAAT GTGCCCAACT ACGCCAATTG TTTTGCCAAC GGCCAACGTT     120

ACGTTCGTAC GCACGTATGT ACCTAGGTAC TTACGGACGT GACTACGGAC ACTTCCGTAC     180

GTACGTACGT TTACGTACCC ATCCCAACGT AACCACAGTG TGGTCGCAGT GTCCCAGTGT     240

ACACAGACTG CCAGACATTC TTCACAGACA CCCC ATG ACA CCA CCT GAA CGT         292
                                      Met Thr Pro Pro Glu Arg
                                      -34             -30

CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC CTT CTG       340
Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu Leu Leu
            -25                 -20                 -15

GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAT GTGAGGCAGC AGGAGAATGG     393
Gly Leu Leu Leu Val Leu Leu Pro Gly Ala His
        -10                 -5
```



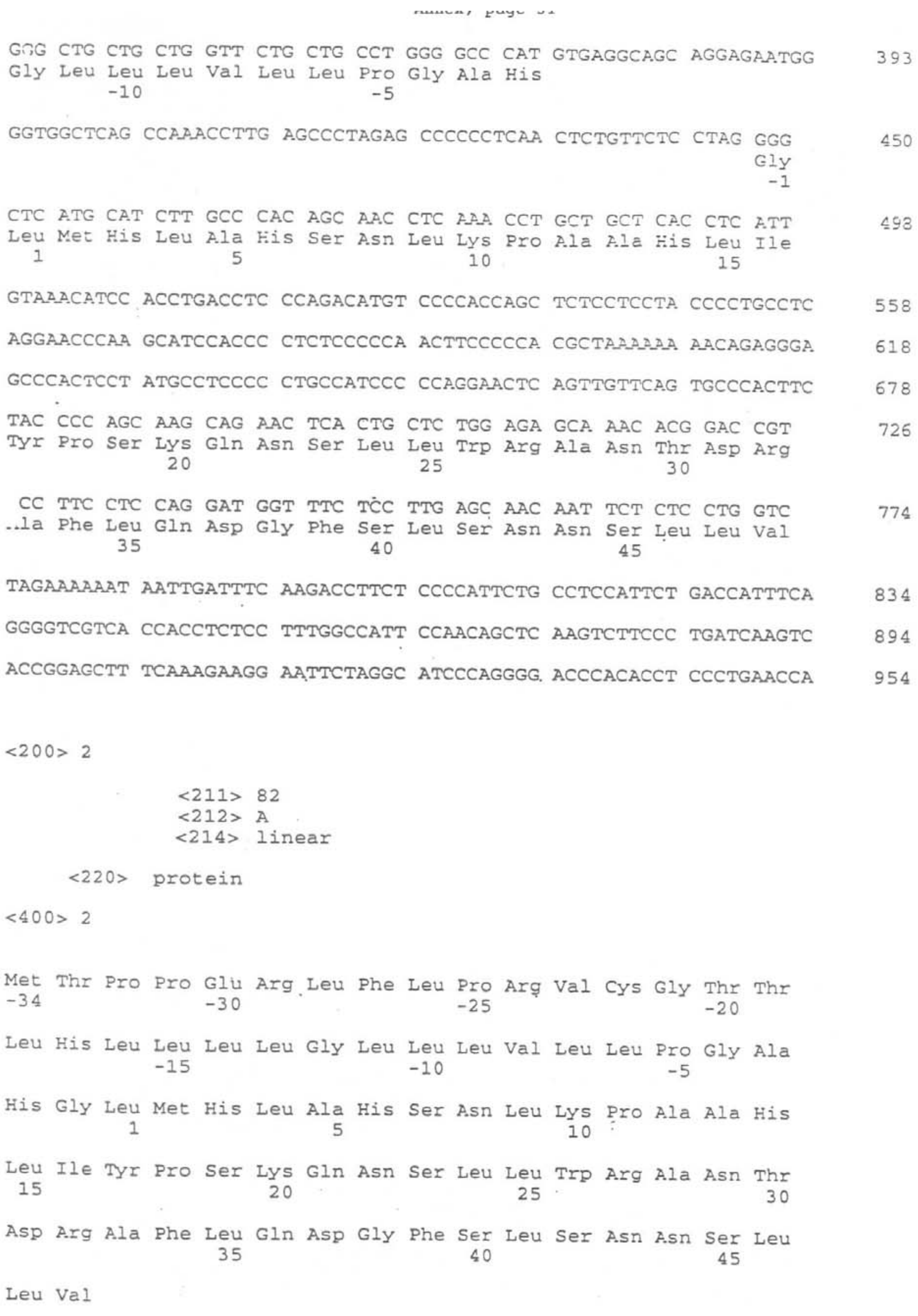
```
GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA CTCTGTTCTC CTAG GGG        450
                                                             Gly
                                                              -1

CTC ATG CAT CTT GCC CAC AGC AAC CTC AAA CCT GCT GCT CAC CTC ATT        498
Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His Leu Ile
  1               5                  10                  15

GTAAACATCC ACCTGACCTC CCAGACATGT CCCCACCAGC TCTCCTCCTA CCCCTGCCTC     558

AGGAACCCAA GCATCCACCC CTCTCCCCCA ACTTCCCCCA CGCTAAAAAA AACAGAGGGA     618

GCCCACTCCT ATGCCTCCCC CTGCCATCCC CCAGGAACTC AGTTGTTCAG TGCCCACTTC     678

TAC CCC AGC AAG CAG AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT        726
Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg
             20                  25                  30

GCC TTC CTC CAG GAT GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC        774
Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val
         35                  40                  45

TAGAAAAAAT AATTGATTTC AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA     834

GGGGTCGTCA CCACCTCTCC TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC     894

ACCGGAGCTT TCAAAGAAGG AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA     954


<200> 2

              <211> 82
              <212> A
<400> 2

Met Thr Pro Pro Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr
-34             -30                 -25                 -20

Leu His Leu Leu Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala
            -15                 -10                  -5

His Gly Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His
          1               5                  10

Leu Ile Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr
 15                  20                  25                  30

Asp Arg Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu
                 35                  40                  45

Leu Val
```



Annex 4 /iv

Sequence listing in accordance with Trilateral proposal now including several optional "language dependent" elements not yet included in the standard and/or feature table.

```
            <111> Deposant
            <112> 28, Rue de St. Petersbourg
            <113> Paris
            <115> France
            <116> 75008

       <130> 2

            <151> Floppy disk
            <152> IBM PC Compatible
            <153> PC-DOS/MS-DOS
            <154> PatentIn Release #1.0, Version #1.25 (OEB)

             <161> EP94123456
             <162> 28-FEB-1994

             <171> FR93123456
             <172> 01-MAR-1993

             <183> 01-0001

<200> 1

             <211> 954
             <212> N
|           <213>  double
|           <214>  linear

|           <220>  genomic DNA

            <291>  CDS
            <292>  1...72

            <291>  misc_feature
            <292>  5...10

<400> 1

ATCGGGATAG TACTGGTCAA GACCGGTGGA CACCGGTTAA CCCCGGTTAA GTACCGGTTA       60

TAGGCCATTT CAGGCCAAAT GTGCCCAACT ACGCCAATTG TTTTGCCAAC GGCCAACGTT      120

ACGTTCGTAC GCACGTATGT ACCTAGGTAC TTACGGACGT GACTACGGAC ACTTCCGTAC      180

GTACGTACGT TTACGTACCC ATCCCAACGT AACCACAGTG TGGTCGCAGT GTCCCAGTGT      240

ACACAGACTG CCAGACATTC TTCACAGACA CCCC ATG ACA CCA CCT GAA CGT          292
                                      Met Thr Pro Pro Glu Arg
                                      -34             -30

CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC CTT CTG        340
Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu Leu Leu
            -25                 -20                 -15
```



Annex, page 31

```
GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAT GTGAGGCAGC AGGAGAATGG      393
Gly Leu Leu Leu Val Leu Leu Pro Gly Ala His
        -10                  -5

GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA CTCTGTTCTC CTAG GGG       450
                                                            Gly
                                                             -1

CTC ATG CAT CTT GCC CAC AGC AAC CTC AAA CCT GCT GCT CAC CTC ATT       498
Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His Leu Ile
  1               5                  10                  15

GTAAACATCC ACCTGACCTC CCAGACATGT CCCCACCAGC TCTCCTCCTA CCCCTGCCTC      558

AGGAACCCAA GCATCCACCC CTCTCCCCCA ACTTCCCCCA CGCTAAAAAA AACAGAGGGA      618

GCCCACTCCT ATGCCTCCCC CTGCCATCCC CCAGGAACTC AGTTGTTCAG TGCCCACTTC      678

TAC CCC AGC AAG CAG AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT       726
Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg
            20                  25                  30

 CC TTC CTC CAG GAT GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC       774
..la Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val
        35                  40                  45

TAGAAAAAAT AATTGATTTC AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA      834

GGGGTCGTCA CCACCTCTCC TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC      894

ACCGGAGCTT TCAAAGAAGG AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA      954
```

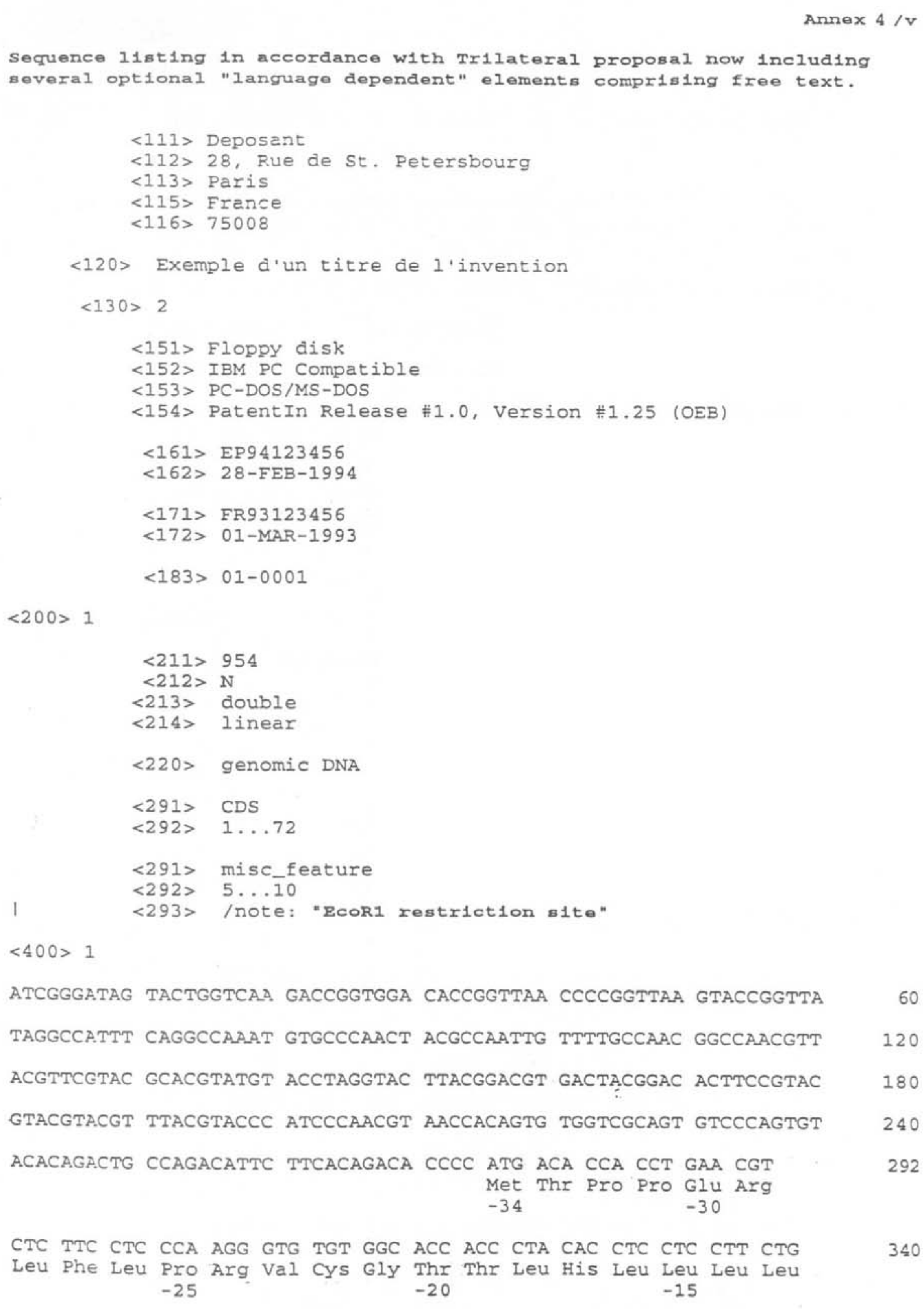
<200> 2

<211> 82
<212> A
<214> linear

<220> protein

<400> 2

```
Met Thr Pro Pro Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr
-34             -30                 -25                 -20

Leu His Leu Leu Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala
            -15                 -10                  -5

His Gly Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His
          1               5                  10

Leu Ile Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr
 15                  20                  25                  30

Asp Arg Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu
                 35                  40                  45

Leu Val
```

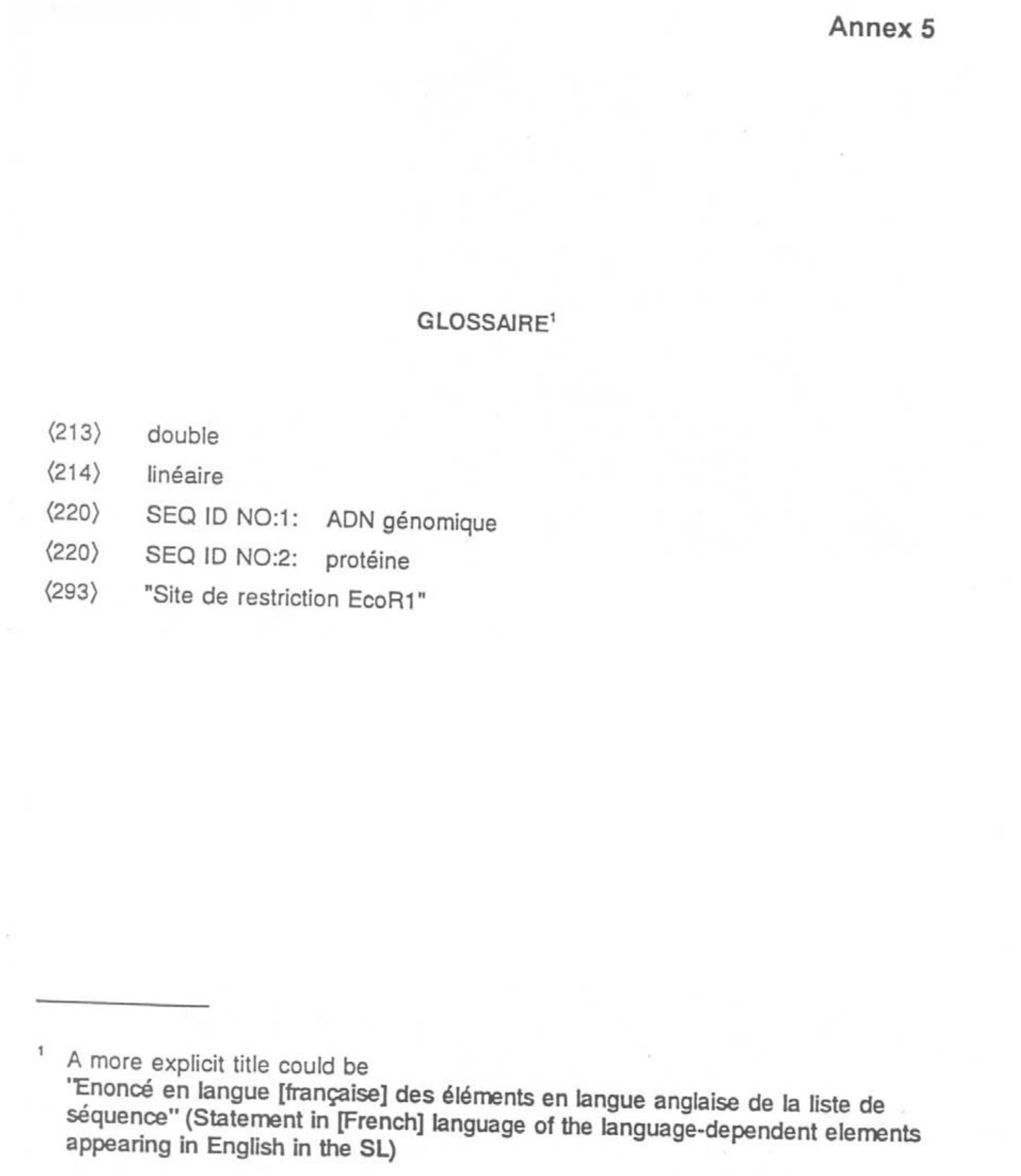


Annex 4 /v

Sequence listing in accordance with Trilateral proposal now including several optional "language dependent" elements comprising free text.

```
          <111> Deposant
          <112> 28, Rue de St. Petersbourg
          <113> Paris
          <115> France
          <116> 75008

     <120>  Exemple d'un titre de l'invention

      <130> 2

          <151> Floppy disk
          <152> IBM PC Compatible
          <153> PC-DOS/MS-DOS
          <154> PatentIn Release #1.0, Version #1.25 (OEB)

           <161> EP94123456
           <162> 28-FEB-1994

           <171> FR93123456
           <172> 01-MAR-1993

           <183> 01-0001

<200> 1

           <211> 954
           <212> N
          <213>  double
          <214>  linear

          <220>  genomic DNA

          <291>  CDS
          <292>  1...72

          <291>  misc_feature
          <292>  5...10
          <293>  /note: "EcoR1 restriction site"

<400> 1

ATCGGGATAG TACTGGTCAA GACCGGTGGA CACCGGTTAA CCCCGGTTAA GTACCGGTTA      60

TAGGCCATTT CAGGCCAAAT GTGCCCAACT ACGCCAATTG TTTTGCCAAC GGCCAACGTT     120

ACGTTCGTAC GCACGTATGT ACCTAGGTAC TTACGGACGT GACTACGGAC ACTTCCGTAC     180

GTACGTACGT TTACGTACCC ATCCCAACGT AACCACAGTG TGGTCGCAGT GTCCCAGTGT     240

ACACAGACTG CCAGACATTC TTCACAGACA CCCC ATG ACA CCA CCT GAA CGT         292
                                      Met Thr Pro Pro Glu Arg
                                      -34                 -30

CTC TTC CTC CCA AGG GTG TGT GGC ACC ACC CTA CAC CTC CTC CTT CTG       340
Leu Phe Leu Pro Arg Val Cys Gly Thr Thr Leu His Leu Leu Leu Leu
            -25                 -20                 -15
```

```
GGG CTG CTG CTG GTT CTG CTG CCT GGG GCC CAT GTGAGGCAGC AGGAGAATGG      393
Gly Leu Leu Leu Val Leu Leu Pro Gly Ala His
        -10                  -5

GGTGGCTCAG CCAAACCTTG AGCCCTAGAG CCCCCCTCAA CTCTGTTCTC CTAG GGG        450
                                                            Gly
                                                             -1

CTC ATG CAT CTT GCC CAC AGC AAC CTC AAA CCT GCT GCT CAC CTC ATT        498
Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His Leu Ile
  1               5                  10                  15

GTAAACATCC ACCTGACCTC CCAGACATGT CCCCACCAGC TCTCCTCCTA CCCCTGCCTC      558

AGGAACCCAA GCATCCACCC CTCTCCCCCA ACTTCCCCCA CGCTAAAAAA AACAGAGGGA      618

GCCCACTCCT ATGCCTCCCC CTGCCATCCC CCAGGAACTC AGTTGTTCAG TGCCCACTTC      678

TAC CCC AGC AAG CAG AAC TCA CTG CTC TGG AGA GCA AAC ACG GAC CGT        726
Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr Asp Arg
             20                  25                  30

 CC TTC CTC CAG GAT GGT TTC TCC TTG AGC AAC AAT TCT CTC CTG GTC        774
Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu Leu Val
         35                  40                  45

TAGAAAAAAT AATTGATTTC AAGACCTTCT CCCCATTCTG CCTCCATTCT GACCATTTCA      834

GGGGTCGTCA CCACCTCTCC TTTGGCCATT CCAACAGCTC AAGTCTTCCC TGATCAAGTC      894

ACCGGAGCTT TCAAAGAAGG AATTCTAGGC ATCCCAGGGG ACCCACACCT CCCTGAACCA      954


<200> 2

              <211> 82
              <212> A
              <214> linear

     <220>  protein

<400> 2


Met Thr Pro Pro Glu Arg Leu Phe Leu Pro Arg Val Cys Gly Thr Thr
-34             -30                 -25                 -20

Leu His Leu Leu Leu Leu Gly Leu Leu Leu Val Leu Leu Pro Gly Ala
            -15                 -10                  -5

His Gly Leu Met His Leu Ala His Ser Asn Leu Lys Pro Ala Ala His
         1               5                  10

Leu Ile Tyr Pro Ser Lys Gln Asn Ser Leu Leu Trp Arg Ala Asn Thr
 15                 20                  25                  30

Asp Arg Ala Phe Leu Gln Asp Gly Phe Ser Leu Ser Asn Asn Ser Leu
                35                  40                  45

Leu Val
```



Annex 5

GLOSSAIRE[1]

⟨213⟩	double
⟨214⟩	linéaire
⟨220⟩	SEQ ID NO:1: ADN génomique
⟨220⟩	SEQ ID NO:2: protéine
⟨293⟩	"Site de restriction EcoR1"

[1] A more explicit title could be "Enoncé en langue [française] des éléments en langue anglaise de la liste de séquence" (Statement in [French] language of the language-dependent elements appearing in English in the SL)

[End of Annex and of document]