

ICS 07.080  
A 40



# 中华人民共和国国家标准

GB/T 34798—2017

---

## 核酸数据库序列格式规范

Formats specifications of nucleotide sequence database

2017-11-01 发布

2018-05-01 实施

中华人民共和国国家质量监督检验检疫总局  
中国国家标准化管理委员会 发布



## 目 次

前言 .....	Ⅲ
1 范围 .....	1
2 规范性引用文件 .....	1
3 术语和定义 .....	1
4 缩略语 .....	2
5 核酸序列格式规范制定的总则 .....	2
6 核酸序列描述规范 .....	2
7 核酸序列特征描述规范 .....	3
8 核酸序列格式规范 .....	5
9 核酸序列文件整体格式规范 .....	5
附录 A (资料性附录) 核苷酸含义表 .....	6
附录 B (资料性附录) 与核酸相关的特征关键词表 .....	7
附录 C (资料性附录) 密码子表 .....	10
附录 D (资料性附录) 修饰碱基表 .....	11
附录 E (资料性附录) 限定词中英文对照表 .....	13
附录 F (资料性附录) 核酸序列文件样例 .....	14
附录 G (资料性附录) 行首大写字母含义表 .....	15
参考文献 .....	16



## 前 言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由全国生化检测标准化技术委员会(SAC/TC 387)提出并归口。

本标准起草单位:深圳华大基因研究院、深圳华大基因科技有限公司、广东省标准化研究院、广东产品质量监督检验研究院。

本标准主要起草人:魏晓锋、陈凤珍、刘克、杜佳婷、李倩一、沈维燕、李启沅、谢强、王娟、谭嘉力、宋祚锟、黄江勇。



## 核酸数据库序列格式规范

### 1 范围

本标准规定了核酸数据库的序列格式,包括生物体基因组核酸序列特征规范制定的总则、序列描述格式规范、序列特征描述规范和序列格式规范等。

本标准适用于生物体基因组核酸数据库序列文件的编写。

### 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 29859 生物信息学术语

ZC 0003 核苷酸和/或氨基酸序列表和序列表电子文件标准

### 3 术语和定义

GB/T 29859 界定的以及下列术语和定义适用于本文件。

#### 3.1

**核酸数据库 nucleic acid database**

以核酸序列为基本内容,并附有核酸序列注释信息的数据库。

#### 3.2

**编码序列 coding sequence**

编码一段蛋白产物的序列,始于起始密码子,终于终止密码子。

#### 3.3

**序列组装 sequence assembly**

基因组长序列打断之后形成较短的序列,通过算法和计算机的帮助,把这些短的序列组装起来成为一条完整有序的序列的过程。

#### 3.4

**甲基化 methylation**

蛋白质和核酸的一种重要的修饰,调节基因的表达和关闭。

#### 3.5

**识别码 identifier**

某个体系中相对唯一的编码。

#### 3.6

**位置 location**

一个或一段碱基在另一段较长碱基上的相对坐标位置。

#### 3.7

**特征限定词 feature qualifier**

用来进一步描述序列的某一类特征的词。



GB/T 34798—2017

### 3.8

#### 修饰碱基 modified base

核酸中主要碱基(腺嘌呤、鸟嘌呤、尿嘧啶、胞嘧啶等)的修饰化合物,核酸转录之后经甲基化、乙酰化、氢化、氟化以及硫化而成,多半是主要碱基的甲基衍生物。

## 4 缩略语

下列缩略语适用于本文件。

CDS:编码序列(coding sequence)

DDBJ:日本核酸数据库(DNA data bank of Japan)

EMBL:欧洲分子生物学实验室(european molecular biology laboratory)

HIV:人类免疫缺陷病毒(human immunodeficiency virus)

ID:识别码(identifier)

Medline:医学文献资料库(medlars on line)

NCBI:美国国立生物技术信息中心(national center for biotechnology information)

RNA:核糖核酸(ribonucleic acid)

UTR:非翻译区(untranslated regions)

## 5 核酸序列格式规范制定的总则

5.1 核酸序列文件应能够与 NCBI、EMBL、DDBJ 等数据库进行共享。

5.2 核酸序列特征描述具有准确性、清晰性、简洁性和明确性,参见 GB/T 29859。

5.3 核酸序列特征内容具有实用性。

## 6 核酸序列描述规范

### 6.1 序列名称

序列名称应符合以下要求:

- 序列名称应为简短的序列描述,包含序列的物种名、基因或蛋白名称及序列功能的简单描述;
- 序列的物种名称命名参考林奈的《自然系统》<sup>[1]</sup>一书中的生物学命名方式;
- 除人类免疫缺陷病毒可用 HIV1 和 HIV2 表示,其他种属应给出属和种的全名,不宜使用通用名如(human)或属名缩写(如代表 Homo sapiens 的 H.sapiens)。

### 6.2 序列编号

序列编号应保证一个序列号码对应一个核酸序列,具有唯一性。序列编号由两个字母加下划线加 6 个数字组成,DNA 序列编号两个字母为 NT(如 NT\_123456),RNA 序列字母为 NM(如 NM\_123456),蛋白序列字母为 NP(如 NP\_123456),整个染色体、质粒等的基因组序列为 NC(如 NC\_123456)。提交一个新的序列会系统产生一个新的序列编号,为保证序列的唯一性,当提交的序列在数据库中已经存在,序列将不能被提交。

### 6.3 序列版本号

序列的版本号是由序列编号加一个点号加版本号(如序列编号.版本号,NM\_123456.1),当一个序列改变,相应的版本号加 1。



#### 6.4 序列长度

序列的长度宜大于 50 bp,无最大值限制。

#### 6.5 日期

日期应为序列最后被公开的日期,此信息只供用户参考,不具有法律保证,不能作为仲裁的判据,不能用来作为优先权声明或专利权请求的依据。日期的格式为 dd-mm-yyyy 格式(如 15-06-1991)。

#### 6.6 碱基总数

碱基总数应为出现在序列中碱基数目的总和,包括 A、C、T、G、U 等碱基数之和,具体核苷酸含义表参见表 A.1。

#### 6.7 分子类型

序列应注明分子类型,分子类型包括 DNA 和 RNA 两种类型。

#### 6.8 测序类型

序列应注明测序的仪器类型。

#### 6.9 组装软件及版本号

序列应注明序列组装所使用的软件。格式为软件名称加版本号,若只有一个版本,版本号可缺省。

#### 6.10 序列参考文献

序列参考文献要求包括:

- a) 每个核酸序列记录要求至少有一篇包含该序列数据的参考文献,如果是已经发表,宜有一个唯一识别码,如 Medline 识别码等;宜提供指向文章数据库的链接,如果未发表,则标识为 Unpublished。
- b) 参考文献包含文献的标题,应为引用文献的标题全名;包含文献作者,应为引文的全部作者名称;以及包含发表的杂志名称、卷、期、页码、年号,如 Yeast 10(11),1503-1509(1994)。
- c) 若引用的参考文献为书本,应包括书本编辑的名称、书的题目、引用的页码,出版者名称、年份信息。

### 7 核酸序列特征描述规范

#### 7.1 关键特征

序列的关键特征需满足以下要求:

- a) 一个序列特征可包含多个关键特征,如 CDS、gene 等,与核酸相关的特征关键词表参见表 B.1;每个关键特征包含位置和限定词两部分;
- b) 核酸序列关键特征词的定義和分类按照 ZC 0003 的规定执行。

#### 7.2 序列位置描述

序列的位置描述类型包括:

- a) 单个碱基,如 23,表示第 23 碱基;
- b) 一个连续的碱基序列,第一个和最后一个碱基用两个点号分开,如 23..79,表示从 23 和 79 之

## GB/T 34798—2017

间的碱基序列。若为互补链,需要在碱基位置前面加 complement,如 complement(3300..4037);若为 5'端部分序列,需要在前面加“<”(如 CDS<1..206);若为 3'端,需要在后面加“>”(如 CDS435..915>);

- c) 两个碱基之间的一个位点,用  $n \sim n+1$  表示,两个碱基之间用尖括号分开,如核酸内切酶位点,23~24,表示第 23 个和第 24 个碱基之间的一个位点;对于一个环形的分子来说,用  $n-1$  表示,其中  $n$  表示分子的总长度,如 1 000-1,表示环形分子总长度为 1 000;
- d) 从一系列碱基选出的单个碱基,第一个碱基和最后一个碱基用点号分开,如 23.79,表示被选的单个碱基在第 23 和第 79 碱基之间;
- e) 多个不连续的序列,用 join 连接,表示为 join(位置 1,位置 2...位置  $n$ ),如 join(23..79,100..160,200..245)。

## 7.3 特征限定词

## 7.3.1 特征限定词表示方法和类型

7.3.1.1 限定词的命名可以包含大写字母(A~Z)、小写字母(a~z)和数字(0~9)、下划线(\_)、连字符(-)、单引号或撇号(')、星号(\*)。

7.3.1.2 序列特征限定词提供了序列特征额外的信息,可以用“=”给限定词赋值,如 note=“text”,限定词包含的信息类型有:

- a) 文本,文本信息应用双引号标记;
- b) 引用,引用的数字宜用方括号“[]”与其他数字区别开来;
- c) 序列,序列应该用双引号标记,如“atgcatt”。

## 7.3.2 限定词定义和分类

## 7.3.2.1 限定词的定义

限定词的定义包括:

- a) 反密码子:tRNA 分子二级结构的反密码环中部的三个相邻的与 mRNA 上的密码子互补配对碱基;
- b) 密码子:RNA 分子中每相邻的三个核苷酸编成一组,在蛋白质合成时,代表某一种氨基酸,密码子表参见表 C.1;
- c) 交叉引用数据库,交叉引用数据库应为支持该核苷酸序列的其他数据库资源,交叉引用数据库宜包含数据库的名称以及交叉引用识别码,数据库和识别码中间用“:”隔开,如 BioProject:PRJNA177352,其中 BioProject 为数据库名称,PRJNA177352 为识别码。若引用一个数据库的多个识别码,直接并排引用多个识别码。如 BioProject:PRJNA174162,PRJNA999998;
- d) 方向:DNA 复制的方向;
- e) 频率:某一特征发生的频率;
- f) 修饰碱基:在 ATGC 四种的不同部位甲基化或进行其他的化学修饰而形成的衍生物,主要修饰碱基及其简写参见表 D.1;
- g) 遗传元素编号:主要指外显子或内含子从 5'到 3'编号,如 number=2 表示第 2 个外显子;
- h) 产物:序列编码的产物名称。

## 7.3.2.2 限定词中英文对照表

限定词中英文对照表参见表 E.1。

## 8 核酸序列格式规范

核苷酸序列宜有开始和结束标志。序列以 ORIGIN 开头,序列在 ORIGIN 的下一行,只包含序列数据。序列以“//”结尾,ORIGIN 和“//”单独为一行。序列每行不宜超过 60 个碱基,每 10 个核苷酸碱基后空一格,该行的行首标明本行序列第一个碱基的编号。

示例:

ORIGIN

```
1 gatectecat atacaacggg atctecacct caggtttaga tctacaacac ggaaccattg
61 ccgacatgag acagtttagt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcatttga agcgcctgaa gttctactaa gggtaggata cctctccgt gcaagaccaa
```

//

## 9 核酸序列文件整体格式规范

序列文件每行首部使用相应的大写字母标识核酸的序列描述、序列特征描述、序列信息等,不能超过 16 个字符,即两个 tab 空格,如“SEQ NAME:”;限定词从第 9 个字符位置开始,如核酸序列文件样例中的“CDS”。核酸序列文件样例参见附录 F。

序列描述内容部分从第 17 个字符位置开始,限定词描述部分从第 25 个字符位置开始。

核算序列描述的大写字母标识含义参见表 G.1。

附 录 A  
(资料性附录)  
核苷酸含义表

核苷酸含义见表 A.1。

表 A.1 核苷酸含义

符号	含义	名称的来源
A	A	腺嘌呤
G	G	鸟嘌呤
C	C	胞嘧啶
T	T	胸腺嘧啶
U	U	尿嘧啶
R	g 或 a	嘌呤
Y	t/u 或 c	嘧啶
M	a 或 c	氨基
k	g 或 t/u	酮基
s	g 或 c	强作用 3H 键
w	a 或 t/u	弱作用 2H 键
b	g 或 c 或 t/u	非 a
d	a 或 g 或 t/u	非 c
h	a 或 c 或 t/u	非 g
v	a 或 g 或 c	非 t, 非 u
n	a 或 g 或 c 或 t/u, 未知, 或其他	任何



**附录 B**  
(资料性附录)  
**与核酸相关的特征关键词表**

与核酸相关的特征关键词见表 B.1。

**表 B.1 与核酸相关的特征关键词**

关键词	说 明
allele	相关的个体或菌株含有相同基因的稳定的其他形式,该形式区别于这一位置的现有的序列(和或许其他序列)
attenuator	(1)存在调节转录的终止的 DNA 区域,它控制了一些细菌操纵子的表达;(2)位于启动子和第一个结构基因之间,引起转录的部分终止的序列区段
C_region	免疫球蛋白轻和重链的恒定区,和 T-细胞受体 $\alpha$ 、 $\beta$ 和 $\gamma$ 链;根据特定的链可包括一个或多个外显子
CAAT_signal	CAAT 盒;位于可能参与 RNA 聚合酶结合的真核生物转录单位的起始点的 75bp 上游的保守序列的一部分;共有序列=GG(C 或 T)CAATCT
CDS	编码序列;对应于蛋白质中的氨基酸序列的核苷酸的序列(位置包括终止密码子);特征包括氨基酸概念上的翻译
conflict	在这一位点或区域,单独确定的“相同”序列有所不同
D-loop	置换环;线粒体 DNA 内的一个区域,其中 RNA 的短的序列与 DNA 的一条链配对,代替了这一区域的原始配对 DNA 链;也用于说明在 RecA 蛋白质催化的反应中,侵入的单链替代双链 DNA 的一条链的区域
D-segment	免疫球蛋白重链的多变区,和 T-细胞受体的 $\beta$ 链
enhancer	顺式-作用序列,它增强了(一些)真核生物启动子的作用,并能在任一方向和与启动子相关的任何位置处(上游或下游)起作用
exon	编码剪接 mRNA 部分的基因组区域;可以含有 5'UTR,所有 CDS 和 3'UTR
GC_signal	GC 盒;位于真核生物转录单位起始点上游的保守的富含 GC 区域,可以以多重拷贝或任一方向存在;共有序列=GGGCGG
gene	鉴定为基因的生物学意义的区域,并已经指定名称
iDNA	间插 DNA;通过几种重组中的任何一种能被消除的 DNA
intron	被转录的 DNA 区段,但通过同时剪接位于其两侧的序列(外显子)即可从转录本内部将其除去
J_segment	免疫球蛋白轻链和重链的连接区段,和 T-细胞受体 $\alpha$ 、 $\beta$ 和 $\gamma$ 链
LTR	长的末端重复,在确定序列的两端直接重复的序列,类型典型地见于逆转录病毒中
mat_peptide	成熟的肽或蛋白质的编码序列;翻译后修饰之后成熟的或最终的肽或蛋白质产物的编码序列;位置不包括终止密码子(与相应的 CDS 不同)
misc_binding	不能用任何其他 Binding 关键词(prim_bind 或 protein_bind)表述的与另一个组成成分共价或非共价结合的核酸中的位点

表 B.1 (续)

关键词	说 明
misc_difference	特征序列与记载中存在的有所不同,并且不能用任何其他不同关键词(conflict、unsure、old_sequence、mutation、variation、allele 或 modified_base)表述
misc_feature	不能用任何其他特征关键词表述的具有生物学意义的区域;新的或少见的特征
misc_recomb	任何一般性的、位点特异性的或复制的重组事件的位点,该位点中有不能用其他重组关键词(iDNA 和 virion)或来源关键词的修饰词(/transposon、/proviral)表述的双螺旋 DNA 的断裂和愈合
misc_RNA	不能用其他 RNA 关键词(prim_transcript、precursor_RNA、mRNA、5' clip、3' clip、5' UTR、3' UTR、exon、CDS、sig_peptide、transit_peptide、mat_peptide、intron、polyA_site、rRNA、tRNA、scRNA 和 snRNA)限定的任何转录本或 RNA 产物
misc_signal	含有控制或改变基因功能或表达之信号的任何区域,所述信号不能用其他 signal 关键词(promoter、CAAT_signal、TATA_signal、-35_signal、-10_signal、GC_signal、RBS、polyA_signal、enhancer、attenuator、terminator 和 rep_origin)表述
misc_structure	不能用其他 structure 关键词(stem_loop 和 D-loop)表述的任何二级或三级结构或构象
modified_base	被指示的核苷酸是经修饰的核苷酸
mRNA	信使 RNA;包括 5'非翻译区(5'UTR)、编码序列(CDS,外显子)和 3'非翻译区(3'UTR)
mutation	在此位置处,相关品系的序列中具有突然的、可遗传的变化
N_region	在重排的免疫球蛋白区段之间插入的额外的核苷酸
old_sequence	在此位置处,所表述的序列修改了此序列以前的版本
polyA_signal	聚腺苷酸化之后内切核酸酶裂解 RNA 转录本所必需的识别区域;共有序列=AATAAA
polyA_site	RNA 转录本上的位点,通过转录后聚腺苷酸化该位点将被加上腺嘌呤残基
precursor_RNA	仍不是成熟的 RNA 产物的任何 RNA 种类;可包括 5'剪切区(5'clip)、5'非翻译区(5'UTR)、编码序列(CDS,外显子)、间隔序列(内含子)、3'非翻译区(3'UTR)和 3'剪切区(3'clip)
prim_transcript	初级(最初的,未加工的)转录本;包括 5'剪切区(5'clip)、5'非翻译区(5'UTR)、编码序列(CDS,外显子)、间隔序列(内含子)、3'非翻译区(3'UTR)和 3'剪切区(3'clip)
prim_bind	起始复制、转录或逆转录的非共价的引物结合位点;包括合成的例如 PCR 引物元件的位点
promoter	参与 RNA 聚合酶的结合以启动转录的 DNA 分子区域
protein_bind	核酸上非共价的蛋白质结合位点
RBS	核糖体结合位点
repeat_region	含有重复单位的基因组区域
repeat_unit	单个重复元件
rep_origin	复制起点;复制核酸以得到两个相同拷贝的起始位点
rRNA	成熟的核糖体 RNA;将氨基酸装配成蛋白质的核糖核蛋白颗粒(核糖体)中的 RNA 成分
S_region	免疫球蛋白重链的开关区;它参与重链 DNA 的重排,导致来自相同 B-细胞的不同免疫球蛋白类的表达



表 B.1 (续)

关键词	说 明
Satellite	短的基本重复单位的很多串联重复(相同或相关的);大多数具有的碱基组成或其他性质与基因组的一般水平不同,这使得它们与大部分(主带)的基因组 DNA 分离开来
scRNA	小的细胞质 RNA;几个小的细胞质 RNA 分子中的任何一个存在于真核生物的细胞质和(有时)核中
sig_peptide	信号肽编码序列;被分泌的蛋白质的 N-末端结构域的编码序列;此结构域涉及新生多肽与膜的结合;前导序列
snRNA	小的核 RNA;很多小的 RNA 种类中的任何一个都被局限于核中;几个 snRNA 参与剪接或其他 RNA 加工反应
source	鉴定序列中特定范围的生物来源;此关键词是强制性的;每一项至少要有有一个跨越整个序列的单一来源关键词;每个序列可允许有一个以上的来源关键词
stem_loop	发卡结构;由 RNA 或 DNA 单链的相邻(反向)互补序列之间的碱基——配对形成的双螺旋区域
STS	序列标记位点;表述基因组上作图界标并能通过 PCR 检测的短的,单拷贝 DNA 序列;通过测定 STS 系列的次序即可作出图谱的基因组区域
TATA_signal	TATA 盒;Goldberg-Hogness 盒;在每个真核生物 RNA 聚合酶 II 转录单位起点前约 25 bp 处发现的保守的富含 AT 的七聚体,它可能涉及使酶定位以正确地起始;共有序列=TATA(A 或 T)A(A 或 T)
terminator	或者位于转录本的末端或者与启动子区域相邻的 DNA 序列,该序列可导致 RNA 聚合酶终止转录;也可以是阻抑蛋白的结合位点
transit_peptide	转运肽编码序列;核编码的细胞器蛋白质 N-末端结构域的编码序列;此结构域参与将蛋白质翻译后运送到细胞器中
tRNA	成熟的转移 RNA,小的 RNA 分子(75 个~85 个碱基长),介导核酸序列翻译成氨基酸序列
unsure	作者不能确定此区域的准确序列
V_region	免疫球蛋白轻链和重链的可变区,和 T-细胞受体 $\alpha$ 、 $\beta$ 和 $\gamma$ 链;编码可变的氨基末端部分;可由 V_segment、D_segment、N_region 和 J_segment 组成
V_segment	免疫球蛋白轻链和重链的可变区段,和 T-细胞受体 $\alpha$ 、 $\beta$ 和 $\gamma$ 链;编码大多数可变区(V_region)和前导肽的最后几个氨基酸
variation	含有来自相同基因的稳定突变的相关系列(例如 RFLP、多态性等),在此(和可能其他)位置处所述相同基因与被表述的不同
3'clip	在加工过程中被切下的前体转录本 3'端大部分区域
3'UTR	不被翻译成蛋白质的成熟转录本的 3'末端区域(终止密码子之后)
5'clip	在加工过程中被切下的前体转录本 5'端大部分区域
5'UTR	不被翻译成蛋白质的成熟转录本的 5'末端区域(起始密码子之前)
-10_signal	Pribnow 盒;细菌转录单位起点上游约 10 bp 处的保守区域,它可能参与结合 RNA 聚合酶;共有序列=TatAaT
-35_signal	细菌转录单位起点上游约 35 bp 处的保守六聚体;共有序列=TTGACA[]或 TGTGACA[]

附 录 C  
(资料性附录)  
密 码 子 表

密码子见表 C.1。

表 C.1 密码子

		第二位碱基			
		U	C	A	G
第一位碱基	U	UUU(Phe/F)苯丙氨酸	UCU(Ser/S)丝氨酸	UAU(Tyr/Y)酪氨酸	UGU(Cys/C)半胱氨酸
		UUC(Phe/F)苯丙氨酸	UCC(Ser/S)丝氨酸	UAC(Tyr/Y)酪氨酸	UGC(Cys/C)半胱氨酸
		UUA(Leu/L)亮氨酸	UCA(Ser/S)丝氨酸	UAA 终止	UGA 终止
		UUG(Leu/L)亮氨酸	UCG(Ser/S)丝氨酸	UAG 终止	UGG(Trp/W)色氨酸
	C	CUU(Leu/L)亮氨酸	CCU(Pro/P)脯氨酸	CAU(His/H)组氨酸	CGU(Arg/R)精氨酸
		CUC(Leu/L)亮氨酸	CCC(Pro/P)脯氨酸	CAC(His/H)组氨酸	CGC(Arg/R)精氨酸
		CUA(Leu/L)亮氨酸	CCA(Pro/P)脯氨酸	CAA(Gln/Q)谷氨酰胺	CGA(Arg/R)精氨酸
		CUG(Leu/L)亮氨酸	CCG(Pro/P)脯氨酸	CAG(Gln/Q)谷氨酰胺	CGG(Arg/R)精氨酸
	A	AUU(Ile/I)异亮氨酸	ACU(Thr/T)苏氨酸	AAU(Asn/N)天冬酰胺	AGU(Ser/S)丝氨酸
		AUC(Ile/I)异亮氨酸	ACC(Thr/T)苏氨酸	AAC(Asn/N)天冬酰胺	AGC(Ser/S)丝氨酸
		AUA(Ile/I)异亮氨酸	ACA(Thr/T)苏氨酸	AAA(Lys/K)赖氨酸	AGA(Arg/R)精氨酸
		AUG(Met/M)甲硫氨酸起始	ACG(Thr/T)苏氨酸	AAG(Lys/K)赖氨酸	AGG(Arg/R)精氨酸
	G	GUU(Val/V)缬氨酸	GCU(Ala/A)丙氨酸	GAU(Asp/D)天冬氨酸	GGU(Gly/G)甘氨酸
		GUC(Val/V)缬氨酸	GCC(Ala/A)丙氨酸	GAC(Asp/D)天冬氨酸	GGC(Gly/G)甘氨酸
		GUA(Val/V)缬氨酸	GCA(Ala/A)丙氨酸	GAA(Glu/E)谷氨酸	GGA(Gly/G)甘氨酸
		GUG(Val/V)缬氨酸	GCG(Ala/A)丙氨酸	GAG(Glu/E)谷氨酸	GGG(Gly/G)甘氨酸

**附 录 D**  
(资料性附录)  
**修饰碱基表**

修饰碱基见表 D.1。

**表 D.1 修饰碱基**

符号	含义
ac4c	4-乙酰胞苷
chm5u	5-(羧羟甲基)尿苷
cm	2'-O-甲基胞苷
cmnm5s2u	5-羧甲基氨基甲基-2-硫代尿苷
cmnm5u	5-羧甲基氨基甲基尿苷
d	二氢尿苷
fm	2'-O-甲基假尿苷
gal q	$\beta$ -D-半乳糖 Q 核苷
gm	2'-O-甲基鸟苷
i	肌苷
i6a	N6-异戊烯基腺苷
mla	1-甲基腺苷
mlf	1-甲基假尿苷
mlg	1-甲基腺苷
mli	1-甲基肌苷
m22g	2'-2-二甲基腺苷
m2a	2-甲基腺苷
m2g	2-甲基鸟苷
m3c	3-甲基胞苷
m5c	5-甲基胞苷
m6a	N6-甲基腺苷
m7g	7-甲基鸟苷
mam5u	5-甲基氨基甲基尿苷
mam5s2u	5-甲氧基氨基甲基-2-硫代尿苷
man q	$\beta$ -D-甘露糖 Q 核苷
mcm5s2u	5-甲氧基羰基甲基-2-硫代尿苷
mcm5u	5-甲氧基羰基甲基尿苷
mo5u	5-甲氧基尿苷
ms2i6a	2-硫代甲基-N6-异戊烯基腺苷

表 D.1 (续)

符号	含义
ms2t6a	N-[(9-β-D-呋喃核糖基-2-硫代甲基嘌呤-6-Y1)氨基甲酰]苏氨酸
mt6a	N-[(9-β-D-呋喃核糖嘌呤-6-Y1)N-甲基氨基甲酰]苏氨酸
mv	尿苷-5-氧化乙酸-甲基酯
o5u	尿苷-5-氧化乙酸
osyw	怀丁氧苷(wybutoxosine)
p	假尿苷
q	Q 核苷
s2c	2-硫代胞苷
s2t	5-甲基-2 硫代尿苷
s2u	2-硫代尿苷
s4u	4-硫代尿苷
t	5-甲基尿苷
t6a	N6-苏氨酰基氨基甲酰基腺苷
tm	2'-O-甲基-5-甲基尿苷
um	2'-O-甲基尿苷
yw	怀丁苷(wybutosine)
x	3-(3-氨基-3-羧基-丙基)尿苷,(acp3)u



附 录 E  
(资料性附录)  
限定词中英文对照表

限定词中英文对照见表 E.1。

表 E.1 限定词中英文对照

中文	英文
反密码子	Anticodon
部分约束	bound_moiety
引用	citation
密码子	codon
起始密码子	codon_start
内含子剪切位点	cons_splice
数据库交叉引用	db_xref
方向	direction
酶学委员会编号	EC_number
证据	evidence
频率	frequency
功能	function
基因	gene
标签	label
染色体上的位置	map
修饰碱基	mod_base
备注	note
数字	number
组织	tissue
部分	partial
表型	phenotype
产物	product
假基因	pseudo
重复序列家族	rpt_family
重复序列的类型	rpt_type
标准名称	standard_name
额外翻译	transl_except
翻译	translation
类型	type

**附 录 F**  
(资料性附录)  
核酸序列文件样例

```

SEQ NAME      Saccharomyces cerevisiae TCP1-beta gene,partial cds,and Axl2p
               (AXL2) and Rev7p (REV7) genes,complete cds.
ACCESSION     NT49845
VERSION       NT49845.1
DATE          21-06-1999
TOTAL BASE    5028bp
MOLECULE      DNA
SEQ METHOD     Illumina
ASSE PROG     SOAPdenovo
ORGANISM      Saccharomyces cerevisiae
REFERENCE     1
AUTHORS       Torpey,L.E.,Gibbs,P.E.,Nelson,J.and Lawrence,C.W.
TITLE         Cloning and sequence of REV7,a gene whose function is required for
               DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL       Yeast 10(11),1503-1509(1994)
PUBMED        7871890
FEATURES      Location/Qualifiers
               CDS
                 (1..206
                 codon_start=3
                 product="TCP1-beta"
                 protein_id="AAA98665.1"
                 db_xref="GI:1293614"
                 translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASE
                 AAEVLLRVDNIIRARPRTANRQHM"
               gene
                 687..3158
                 gene="AXL2"

```

**ORIGIN**

```

1 gatctccat atacaacggt atctccact caggittaga tctcaacaac ggaaccattg
61 cggacatgag acagttaggt atcgctgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcacttga agcgcgtgaa gttctactaa ggggtggataa catcctcgt gcaagaccaa
181 gaaccgcaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaacg
241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatecacta tataattcaa
...
4 801 gatetcaagt tatitggagtc ttacgccaat tgetttgtat cagacaattg actctetaac
4 861 ttctecactt caetgtcgag ttgctcggtt ttacgggaca aagatttaac ctggttttct
4 921 ttttcagtgt tagattgtct taattctttg agctgttctc ttagctctct atatttttct
4 981 tgccatgact cagattctaa tttaagcta ttcaatttct ctttgatc

```

//



附 录 G  
(资料性附录)  
行首大写字母含义表

行首大写字母含义见表 G.1。

表 G.1 行首大写字母含义

英文	中文
SEQ NAME	序列名称
ACCESSION	序列编号
VERSION	版本号
DATE	日期
TOTAL BASE	总碱基数
MOLECULE	分子类型
SEQ METHOD	测序方法
ASSE PROG	组装程序
ORGANISM	物种名称
REFERENCE	参考文献
AUTHORS	作者
TITLE	文献标题
JOURNAL	杂志
PUBMED	文献服务检索系统
FEATURES	关键特征
CDS	编码序列(限定词)
ORIGIN	序列开始标识

GB/T 34798—2017

#### 参 考 文 献

- [1] Linnaeus, Carolus, *Systema naturae per regna trianaturae; secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* 10th. Stockholm: Laurentius Salvius, 1758 (Latin).
-



中 华 人 民 共 和 国  
国 家 标 准  
核酸数据库序列格式规范  
GB/T 34798—2017

\*

中国标准出版社出版发行  
北京市朝阳区和平里西街甲2号(100029)  
北京市西城区三里河北街16号(100045)

网址: [www.spc.org.cn](http://www.spc.org.cn)

服务热线: 400-168-0010

2017年11月第一版

\*

书号: 155066 • 1-57676

版权专有 侵权必究



GB/T 34798-2017