ST.26 标准及配套软件应用的可行性研究报告

中国专利信息中心

信息研究处

2019 年

摘 要

根据标准委员会的要求，2022 年 1 月 1 日起所有知识产权局接收的新申请都要符合 ST.26 标准。为了配合该行动，国际局开发了 ST.26 编著和验证软件工具。在 2022 年 1 月 1 日之前，中国需要对 ST.26 标准的结构、内容和发展变化进行充分研究，对 ST.26 编著和软件工具的主要特点、系统和硬件配置要求以及软件部署方式等问题进行全面评估，以便为 2022 年 1 月 1 日正式启动符合 ST.26 标准的申请做好准备。

本研究通过综合运用文献研究、调查研究、比较研究和专家论证等方法，在对 ST.26 标准和 ST.25 标准进行全面分析和比较的基础上，考察 ST.26 标准相对于 ST.25 标准的变化，以及这些变化对于智能化升级系统设计的影响；在深入分析国际局开发的 ST.26 编著和验证软件工具的组成、特点、功能和使用要求的基础上，评估该工具在智能化升级系统中应用的可行性。

ST.26 标准源于 2010 年欧洲专利局（EPO）提出的建立一个基于 XML 格式的生物序列标准，其目的在于使申请人在专利申请中所撰写的序列表能在国际和国家阶段都被接受；提高序列表述的精确度和质量，从而有利于申请人、公众和审查员更容易传播该序列；有利于序列数据的检索；以及允许序列数据以电子形式进行交换和输入计算机数据库。ST.26 标准于 2016 年在 WIPO 的第四届会议（续会）上正式通过，迄今已发布 4 个版本。

符合 ST.26 标准的序列表包括通用数据部分和序列数据部分。通用数据部分主要包括著录信息，用于与该序列表对应的专利申请进行关联；序列数据部分主要由关于该序列信息的一个或多个数据元素组成，数据元素又包括各种特征关键词和限定词。与 ST.25 标准相比，ST.26 标准的变化主要体现在适用范围、数据格式、标准内容等方面。ST.26 标准的适用范围更加广泛，不仅适用于国际申请，也适用于国内或区域程序的专利申请。ST.26 标准采用了 XML 格式，在数据管理、交换和解析方面更加安全、便捷；ST.26 标准的著录信息和序列信息更加丰富，采用特征关键词和限定词对特征信息进行描述，序列的表述更加详细和专业；此外，ST.26 标准注重与国际标准（ISO 639）、相关产业建议（INSD 等）的一致性。

ST.26 标准与 ST.25 标准之间的差异对智能化升级系统设计的影响，主要体现在应用系统设计、数据库设计和历史数据处理三个方面。在电子申请系统的设计中，需要在客户端及交互式平台中提供符合 ST.26 标准的序列表的生成、导入、编辑和校验功能；在智能审查系统中，需要提供符合 ST.26 标准的序列表的校验、形式审查及展示功能；在智能检索系统中，需要提供代码转换、序列比对和序列翻译功能。在数据库设计方面，需要包含 ST.26 标准的 DTD 规范中的重要数据项，在数据字段的设置、字段长度、属性及其取值与 ST.26 标准的要求相符合。此外，在智能化系统升级的设计中还要考虑历史数据的处理问题。

中国的生物序列目前主要是按照 ST.25 标准执行，使用过程中也存在诸多问题，例如 CEPCT 电子申请客户端不支持生物序列的编辑和校验，CPC 电子申请客户端的序列表验证程序偶尔出现错误，PCT 国际申请中序列表因格式和要求不一致而只能以光盘提交，检索系统缺少序列转换、序列比对和序列翻译等功能。在 2022 年 1 月 1 日来临之前，中国需要从计划制定、规则修改、工具测试、IT 系统升级和人员培训等方面做好准备。

国际局开发的 ST.26 标准配套的软件工具由 WIPO Sequence（申请人编著和验证程序），WIPO Sequence Validator（局端验证程序），以及 WIPO Sequence Server（更新和发布服务器）组成。

WIPO Sequence 能够实现申请端所需的序列创建、导入、编辑、验证、导出等功能，支持符合 ST.25 标准的 txt 文件生成 ST.26 标准的 XML 文件，基本能够满足申请人对生物序列的编著和验证需求。但是，目前的 WIPO Sequence 是独立的应用程序，如何在智能化升级系统的电子申请、智能审查和智能检索系统中集成，实现相应的导入、编辑、展示等功能，还需要技术人员进一步开发。

WIPO Sequence Validator 可以作为微服务部署在服务器环境中，与知识产权局使用的其他业务解决方案的应用程序通信，可以对申请人提交序列表数据提供验证服务。但是，WIPO Sequence Validator 目前仅支持符合 ST.26 标准的 XML 文件格式的生物序列表的验证，验证界面不进行错误提示，也不支持在线编辑。如果验证中存在错误或警告，也仅以报告（XML 格式）

的形式在 report 文件夹中给出。此外，XML 格式不利于人工阅读。因此，在智能化升级系统中集成 WIPO Sequence Validator，还需要考虑如何对 XML 格式的文件进行处理，将序列表以利于人工阅读的形式进行呈现。

关键词：ST.26，ST.25，生物序列，核苷酸，氨基酸，标准，系统设计

目 录

第一章 研究概况

一、 研究背景

2017 年 5 月，在 WIPO 标准委员会（CWS）第五届会议上，CWS 要求所有知识产权局同时从 ST.25 过渡到 ST.26[①]，以国际申请日为准，2022 年 1 月 1 日起各局收到的新申请都必须符合 ST.26 标准。2018 年 10 月，在 CWS 第六届会议上，国际局在其提交的《产权组织标准 ST.26 编著和验证软件工具开发情况的状态报告》中指出 ST.26 编著和验证软件工具的开发项目计划于 2019 年完成。

当前，国家知识产权局正在进行专利审查和检索系统的智能化升级。为了顺利地将 ST.26 编著和软件工具应用到新系统中，前期需要对 ST.26 标准的结构、内容和发展变化进行充分研究，对 ST.26 编著和软件工具的主要特点、系统和硬件配置要求以及软件部署方式等问题进行充分调研。

二、 研究目标

本研究的目标之一是在对 ST.26 标准和 ST.25 标准的全面分析和比较的基础上，考察标准变化对智能化升级系统设计的影响。

本研究的目标之二是在深入分析国际局开发的ST.26编著和验证软件工具的组成、特点、功能和使用要求的基础上，评估该工具在智能化升级系统中应用的可行性。

三、 研究现状

通过检索发现，国内外关于生物序列标准的研究并不多见。除了 WIPO 的生物序列标准制定过程中的讨论、调查和会议资料外，研究论文极其稀少。国外的研究论文主要包括：Osmat A. Jefferson 等（2015）[②]在“全球专利实践中生物序列的公开披露”一文中对美国专利商标局、世界知识产权组织和

① 如无特别说明，本文中 ST.25 是 2009 版的，ST.26 是 2019 版的。

② Osmat A. Jefferson, Deniz Köllhofer, Prabha Ajjikuttira, et al. Public disclosure of biological sequences in global patent practice[J], World Patent Information, 2015 (43):12-24.

欧洲专利局在1990-2013年间对核苷酸或肽序列的申请和公布的格式及可专利性要求的变化。Poliana Belisário Zorzal 等（2019）[①]从法律和实例的角度比较了巴西、欧洲和美国专利局在授权程序中对生物序列的公开充分性和类权利要求的规定。国内的研究论文主要是曲超等（2012）[②]和张春华等（2015）[③]对 WIPO 的 ST.25 标准和 ST.26 标准进行了对比或差异分析。

此外，世界主要知识产权局（以美国和欧洲为例）对专利申请中的生物序列从法律和法规层面做出了规定。

(一) 美国对专利申请中生物序列的规定

1998 年 7 月美国专利商标局 USPTO 公布了对含核苷酸序列或氨基酸的专利申请的要求，以符合国际标准，采用非特定语言格式并使用数字标识符取代当前序列表中的主题行。此时规则被修改为与 WIPO 新标准 ST.25 一致。目前《联邦行政法典》第 37 篇（Title 37-Code of Federal Regulations Patents, Trademarks, and Copyrights；37 CFR）中§1.821-1.825 条规定了对包含核苷酸序列和/或氨基酸序列公开的专利申请相关的要求[④]，内容如下：

§1.821 对专利申请中的核苷酸和/或氨基酸序列披露的规定：对世界知识产权组织标准 ST.25（1998 版），包括附件 II 中的表 1-6，以引用的方式直接并入；核苷酸和/或氨基酸序列的解释为四个及以上不分支氨基酸序列或十个及以上不分支核苷酸序列；包含核苷酸和/或氨基酸序列披露的专利申请，就核苷酸和/或氨基酸序列的呈现和描述方式而言，应完全符合§1.821 至 1.825 的要求；包含核苷酸和/或氨基酸序列的专利申请必须包含纸质或光盘副本（作为披露的单独部分）。

§1.822 条规定了用于核苷酸和/或氨基酸序列数据的符号和格式。其中(b)款中规定：表示核苷酸和/或氨基酸序列字符的代码应符合 WIPO 标准 ST.25（1998 版），并对核苷酸和氨基酸序列的格式表示方法做了详细的说明。

① Poliana Belisário Zorzal, Fabricia Pires Pimenta,Antonio Alberto Ribeiro Fernandes, et al. Sufficiency of disclosure and genus claims for protection of biological sequences: a comparative study among the patent offices in Brazil, Europe and the United States[J]. Biotechnology Research and Innovation, 2019,3(1): 91-102.
② 曲超，张松，曲晓光. WIPO 标准 ST.26 与 ST.25 对比研究分析[J]. 中国标准化, 2012(9):50-52.
③ 张春华，马晓蕾，李茉. WIPO 标准 ST.25 与 ST.26 差异分析[J]. 标准科学, 2015(2):67-71.
④ 美国《联邦行政法典》中有关核苷酸序列和/或氨基酸序列的规定[EB/OL]. [2019-08-08]. https://www.uspto.gov/web/offices/pac/mpep/mpep-9020-appx-r.html#d0e333653.

§1.823 至 1.825 条规定了核苷酸和/或氨基酸序列表，作为专利申请的一部分，以书面形式和光盘形式提交的具体要求；以计算机可读形式提交核苷酸和/或氨基酸序列的形式和格式；以及序列表及其计算机可读副本的修订或替换规则。

美国的生物序列规则的要求在核苷酸和/或氨基酸序列的呈现和描述方式方面完全符合 ST.25 标准，但整体不如 WIPO 的 ST.25 标准要求严格[①]。

(二) 欧洲专利局对专利申请中生物序列的规定

2007 年 12 月欧洲专利局为了适应含有关核苷酸和氨基酸序列的欧洲专利申请的需要，在《欧洲专利公约实施细则》（Implementing Regulations to the Convention on the Grant of European Patents）中加入第 30 条，对于欧洲专利申请中核苷酸和氨基酸序列的要求，第（1）款规定如果欧洲专利申请中披露了核苷酸或氨基酸序列，说明书应包含符合欧洲专利局主席就核苷酸和氨基酸序列的标准化表示制定的规则的序列表。根据欧洲专利局主席 2007 年 7 月 12 日关于序列表归档的决定 C.1, A1(1)，如果在欧洲专利申请中披露了核苷酸或氨基酸序列，说明书应包含符合 WIPO 的 ST.25 标准（1998 版）的序列表[②]。因此，欧洲专利公约实施细则第 30 条（1）款所指的规则即 WIPO 的 ST.25 标准。

2013 年 10 月关于序列表的归档被修改为：“1.1 如果在欧洲专利申请中披露了核苷酸或氨基酸序列，说明书必须包含符合 WIPO 的 ST.25 标准的序列表”[③]。同时，“1.4 根据主席决定 A1(1)，序列表必须以电子形式提交，即文本格式（TXT）。有关文件格式的更多信息见本标准。序列表不得再以纸质形式提交，或者如果是电子形式提交申请，则不得以 PDF 格式提交（见主席决定 A1(1)和(2)）。申请人同时以纸质或者 pdf 格式提交序列表的，应当提交电子版、纸质或者 pdf 格式的序列表相同的声明。在这种情况下，纸张或 PDF 格式将在下一个过程中被忽略。”至此，欧洲专利局对序列表的归

① Osmat A. Jefferson, Deniz Köllhofer, Prabha Ajjikuttira, et al. Public disclosure of biological sequences in global patent practice[J], World Patent Information, 2015 (43):12-24.
② Special edition No. 3, OJ EPO 2007, C.1.
③ 欧洲专利授权程序中关于序列表的规定[EB/OL]. [2019-06-27]. http://archive.epo.org/epo/pubs/oj013/11_13/11_5423.pdf.

档标准进行了强制性的规定。

(三) 生物序列标准使用情况

根据世界知识产权组织标准委员会（CWS）对世界知识产权组织 ST.25 标准的使用情况及未来对 ST.26 标准使用计划的调查[①]，统计整理得到 WIPO 生物序列相关标准的使用情况（详见表 1）。

表 1 WIPO 的生物序列相关标准使用情况

国别	ST.25	ST.26 使用计划
AR	未使用	--
AU	完全使用	--
BA	未使用	无计划
BD	部分使用	--
CA	完全使用	有计划
CH	未作答	未作答
CN	部分使用	有计划
CO	部分使用	--
CZ	完全使用	无计划
DE	完全使用	有计划
EA	完全使用	有计划
EC	未使用	无计划
EM	未使用	无计划
ES	完全使用	有计划
GB	完全使用	有计划
GE	完全使用	无计划
HN	完全使用	无计划
HR	部分使用	无计划
HU	完全使用	--
IL	完全使用	--
IT	部分使用	无计划
JP	完全使用	有计划
KG	完全使用	--
KR	完全使用	有计划
LT	完全使用	--
MD	完全使用	--
MX	部分使用	--
OM	完全使用	--

① WIPO 标准使用情况的调查[EB/OL]. [2019-06-27]. https://www3.wipo.int/confluence/display/usestandards/WIPO+Standard+ST.25%3A+Presentation+of+nucleotide+and+amino+acid+sequence+listings.

RU	完全使用	--
SA	未使用	有计划
SE	完全使用	无计划
SK	完全使用	有计划
SV	未使用	--
TH	完全使用	无计划
TN	部分使用	--
TT	完全使用	--
UA	完全使用	有计划
UG	未使用	--
US	完全使用	--
ZA	未使用	--
注：--表示答复中未明确是否有计划。		

四、 研究内容及思路

(一) 研究内容

ST.26 标准及配套软件应用的可行性研究的目标主要包括两个方面：一是通过深度解析 ST.25 和 ST.26 标准的内容，分析 ST.26 标准相对于 ST.25 标准的变化，以及这些变化对于智能化升级系统设计的影响；二是在深入分析国际局开发的 ST.26 编著和验证软件工具特点和使用要求的基础上，全面评估该工具在智能化升级系统中应用的可行性。在对 ST.26 标准及配套软件进行系统研究的基础上，分析中国生物序列标准现状和使用情况，为 ST.26 标准在中国的实施做好理论储备和规划建议。

围绕上述目标，本研究拟定的研究内容主要包括：

1. ST.25 和 ST.26 标准的全面分析
2. ST.26 和 ST.25 标准的比较
3. 世界主要知识产权局对 ST.26 标准的实施计划
4. 中国生物序列标准现状和使用情况分析
5. ST.26 编著和验证软件工具的深入分析，包括工具组成、特点和功能、工具对系统和硬件的要求以及工具的部署方式等。
6. ST.26 编著和验证软件工具的全面评估，包括 WIPO 软件工具的获取途径、工具的测试、使用 WIPO 软件工具的利弊、自主开发软件工具的利弊等方面。

(二) 研究思路

围绕本研究的目标及拟定的研究内容，本研究的逻辑思路如图 1 所示。

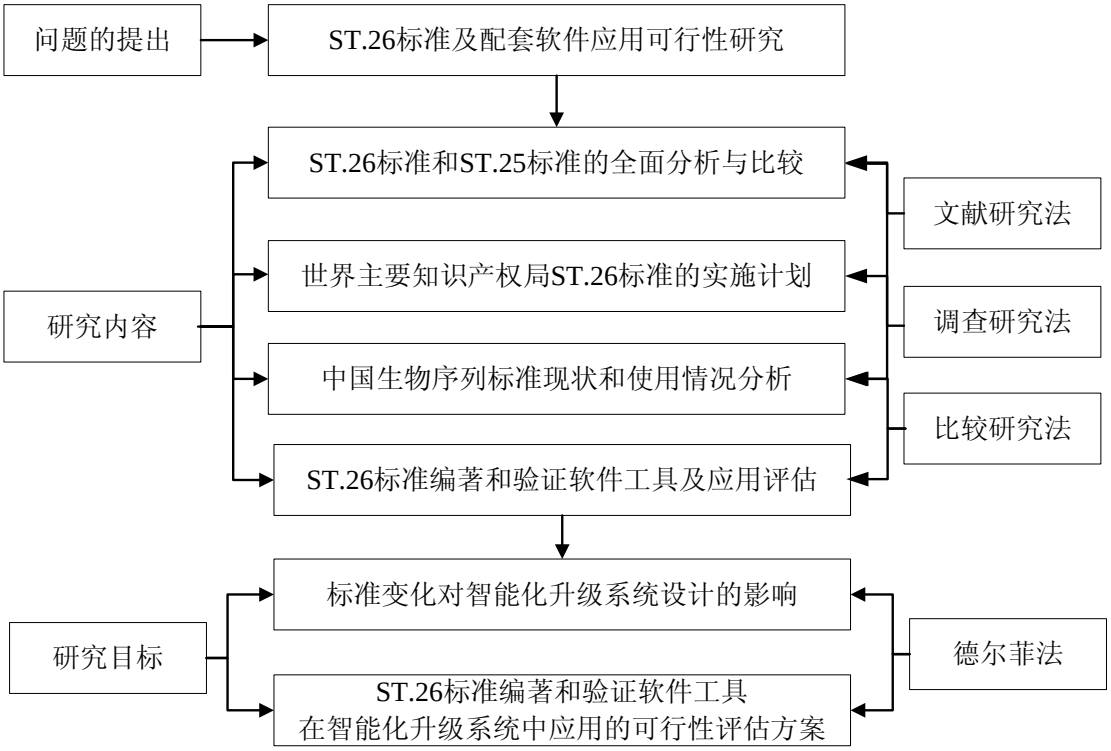


图 1 逻辑思路示意图

根据逻辑思路图，本研究主要从以下 6 章内容展开：

第 1 章为研究概况部分，主要介绍研究背景、研究目标、国内外研究现状、研究内容及思路和研究方法。

第 2 章为 ST.26 标准及其发展历程，主要通过梳理 ST.26 标准的缘起和发展理清 ST.26 标准的前世今生；通过对 ST.26 标准的目标、结构及内容的深入分析，全面了解 ST.26 标准；通过考察世界主要知识产权局的实施计划，提出中国的实施建议。

第 3 章为 ST.26 标准与 ST.25 标准的比较研究，主要是在简要介绍 ST.25 标准的基础上，从标准的适用范围、标准的格式和标准的内容等方面对两个标准进行全面比较，并考虑标准变化可能对智能化升级系统设计的影响。

第 4 章为中国生物序列的现状，主要是在对中国生物序列标准的目的、范围和使用情况分析的基础上，找出中国生物序列存在的问题，并提出解决这些问题的措施。

第 5 章为 ST.26 标准配套软件工具的评估，主要是在对 ST.26 配套软件

工具的组成、特点和功能进行深入分析的基础上，评估其在智能化升级系统中应用的可行性。

第 6 章为总结与展望部分，主要是在对前几章的研究结论进行归纳和总结的基础上，就本研究存在的不足和未来努力的方向进行展望。

五、 研究方法

本研究拟采用的研究方法包括文献研究法、调查研究法、比较研究法和德尔菲法。

(一) 文献研究法

文献研究主要应用于文献综述及对 ST.25 和 ST.26 标准的全面分析阶段。通过文献研究对国内外生物序列标准的现状进行、ST.25 标准和 ST.26 标准进行全面地掌握，从而为后续研究奠定基础。

(二) 调查研究法

调查研究主要应用于对世界主要知识产权局（美、日、欧、韩、国际局）的 ST.26 标准实施计划和中国生物序列标准的使用情况的调查，从而明确中国生物序列标准的现状，并在借鉴国际经验的基础上，提出中国从现有标准向 ST.26 标准过渡的实施建议。

(三) 比较研究法

比较研究法主要应用于对 ST.25 和 ST.26 两个标准的比较，旨在通过对其适用范围、格式和内容的全面比较，明确两个标准之间的差异，并进一步考察标准变化对智能化升级系统设计的影响。

(四) 德尔菲法

德尔菲法主要应用于实施建议和工具的可行性评估报告阶段，旨在充分求证专家意见，以使提出的建议更具有可操作性和可行性评估结论更准确。

第二章 ST.26 标准及其发展历程

一、 ST.26 标准简介

2010 年 9 月，欧洲专利局（EPO）第一次提出讨论建立一个新的基于 XML 格式的生物序列标准 ST.26，即：《Presentation of nucleotide and amino acid sequence listings based on eXtensible Markup Language(XML)，基于 XML 格式的核苷酸和氨基酸序列表的表述标准》；2010 年 10 月，CWS 第一届会议上讨论并采纳了 EPO 的建议，决定设立第 44 号任务[①]；成立由 EPO 牵头的序列表（Sequence Listings，SEQL）工作队负责处理该任务，并就该标准对《PCT 行政规程》附件 C 可能产生的影响与 PCT 相关机构进行联络；要求工作队提交新的 WIPO 标准提案并对标准 ST.25 进行必要的修改，以供 CWS 在 2011 年举行的第二届会议上审议。

(一) ST.26 标准的目标

提出 ST.26 标准的目的在于改变 ST.25 附属于 PCT 行政规程的现状，弥补 ST.25 的缺点，为申请人和各知识产权局提供更加丰富的专利申请信息；打破专利世界特有的 TXT 格式标准，借鉴国际公共数据库的存储模式，为那些真正与生物序列打交道的发明人提供更加专业的生物序列信息。

ST.26 标准的应用目的在于：

- 使申请人在专利申请中所撰写的序列表能在国际和国家阶段都被接受；
- 提高序列表述的精确度和质量，从而有利于申请人、公众和审查员更容易传播该序列；
- 有利于序列数据的检索；
- 以及允许序列数据以电子形式进行交换和输入计算机数据库。

① 第 44 号任务的描述为：“制定一项关于基于可扩展标记语言（XML）的核苷酸和氨基酸序列表表示方法的建议，以作为 WIPO 标准通过。提交这项新 WIPO 标准的提案时，应一并提交报告，说明该标准对现有 WIPO 的 ST.25 标准的影响，包括提出对 ST.25 标准的必要修改”。

(二) ST.26 标准的结构

ST.26 标准由主体文件和七个附件组成，其中主体文件对标准的范围、序列的表示（核苷酸序列、氨基酸序列、特殊情况的表示）、XML 中序列表的结构（根元素、通用信息部分、序列数据部分、特征表、特征关键词、强制性特征关键词、特征位置、特征限定词、强制性特征限定词、限定词元素、自由文本、编码序列和变体）进行了规范。

附件 I——受控词表，主要包括核苷酸表、修饰的核苷酸表、氨基酸表、修饰的氨基酸表、核苷酸序列特征关键词、核苷酸序列特征限定词、氨基酸序列特征关键词、氨基酸序列特征限定词和基因编码表。

附件 II——序列表的文档类型定义（DTD），用于定义 XML 的结构，对元素（子元素）、属性和取值等做出要求。

附件 III——序列表样例（XML 文件）。

附件 IV——Unicode 基本拉丁码表中的字符子集在序列表的 XML 实例中的使用。

附件 V——附加的数据交换要求（仅适用于专利局），规定了与 INSD 成员交换数据的专利局对序列填充元素 INSDSeq_other-seqids 的要求。

附件 VI 及其附录——指导文件，用于指导申请人和知识产权局理解并同意包含和表示序列的公开要求，主要包括序列索引以及附录，附录中包含 XML 中的序列表。

附件 VII——序列表从 ST.25 向 ST.26 转变的建议。

(三) ST.26 标准的序列

ST.26 标准确立了专利申请中公开的序列的核苷酸和氨基酸序列表的呈现要求。符合 ST.26 标准的序列表包括通用数据部分和序列数据部分。通用部分主要包括著录信息，用于与该序列表对应的专利申请进行关联；序列数据部分主要由关于该序列信息的一个或多个数据元素组成，数据元素又包括各种特征关键词和限定词。

符合 ST.26 标准的序列表中的序列是通过枚举残基的方式在专利申请中公开的序列，不包括具有少于 10 个具体定义的核苷酸或少于 4 个具体定义

的氨基酸的序列。

1. 序列的表示

ST.26 中要求每个序列必须被分配独立的序列标识号。序列标识号从 1 开始，按照整数连续递增。当序列标识号没有对应的序列出现时（如“有意跳过的序列”），序列位置应用“000”。序列的总数必须在序列表中指出，且必须等于序列标识号的总数。

对于序列的表示，ST.26 标准中分别对核苷酸序列和氨基酸序列做出了规定。核苷酸序列部分对普通核苷酸序列、双链核苷酸序列、修饰的核苷酸序列和未知的核苷酸序列的表示做了详细说明，氨基酸序列部分对普通氨基酸序列、修饰的氨基酸序列、未知的氨基酸序列和含有空白或内部终止符的氨基酸序列做了详细说明。

在序列的表示部分还对核苷酸和氨基酸序列的序列编号、序列代码和序列中歧义符号的使用做了明确规定。此外，ST.26 标准对于特殊的序列（如变体序列、编码序列、人工源序列、未知源序列、片断组成的序列、缺口分开的序列、以及“n”或“X”残基分开的序列）的表示进行了说明。

2. 序列表的元素

依据 ST.26 标准，序列表必须以附件 II 中 DTD 定义的 XML 文件单独呈现，整个序列表必须包含在一个文件中，且使用 Unicode UTF-8 对文件进行编码。图 2 给出了附件 II 中 DTD 定义的内容模式图。从图 2 可以看出序列表的元素组成和层级结构。

(1) 根元素

在一个符合 ST.26 标准的 XML 文件中，根元素是 ST26SequenceListing，其包括一个强制性属性——dtdVersion（创建序列表文件的 DTD 版本信息），和三个可选属性——filename（序列表的文件名）、softwarename（创建该文件的软件名）、softwareVersion（生成该文件的软件版本）和 productionDate（序列表文件的生成日期）。

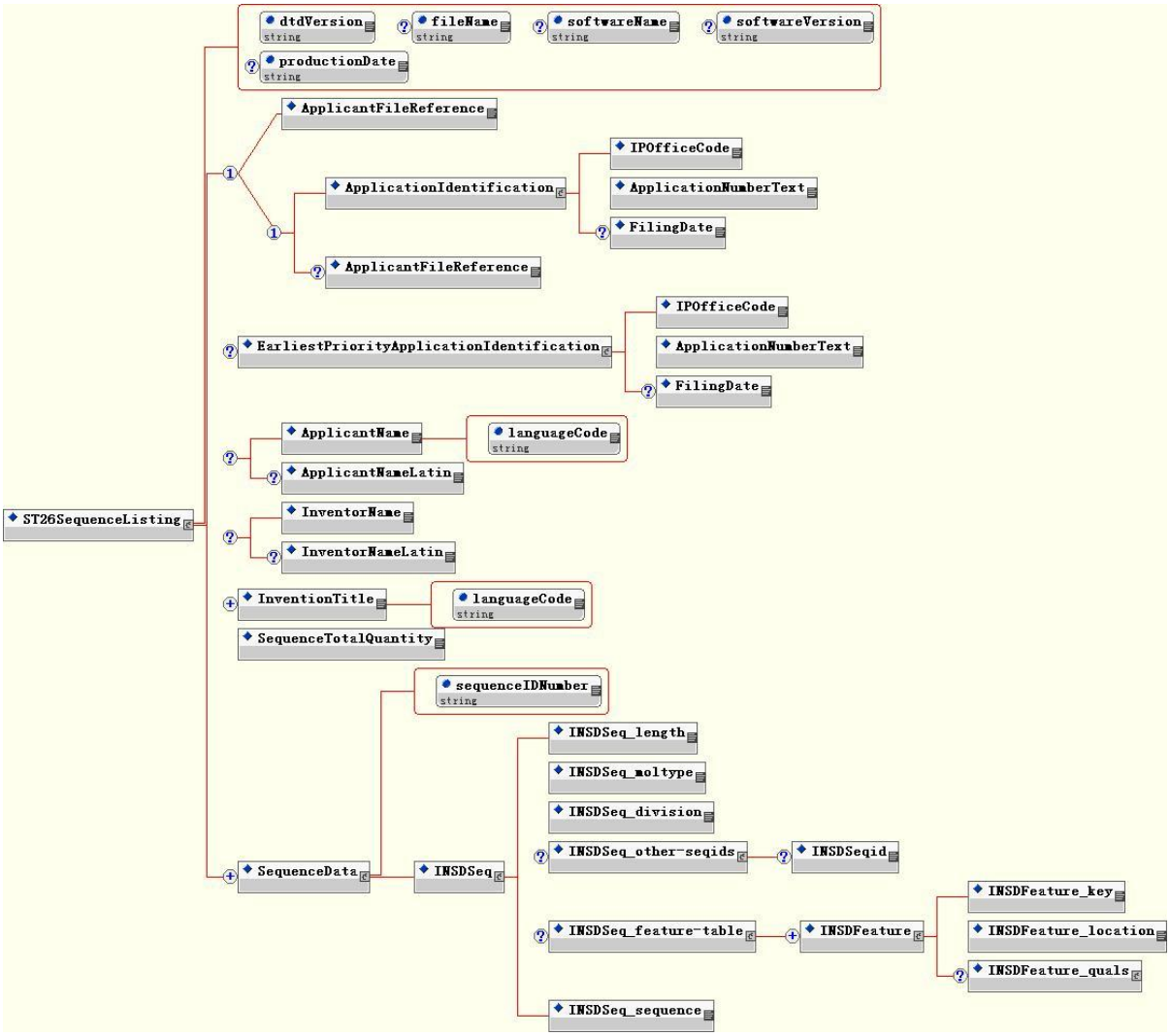


图 2 DTD 定义的内容模式图

(2) 通用信息部分的元素

通用信息部分的元素主要和专利申请信息有关，包括 ApplicationIdentification[申请识别信息，由 IPOfficeCode（知识产权局代码）、ApplicationNumberText（申请号）和 FilingDate（申请日）组成]或 ApplicantFileReference（申请文件参考信息）、EarlistPriorityApplicantionIdentification（最早优先权信息，由 IPOfficeCode（知识产权局代码）、ApplicationNumberText（申请号）和 FilingDate（申请日）组成）、ApplicantName（申请人名称，包括强制性属性 languageCode）、ApplicantNameLatin（申请人的拉丁名称）、InventorName（发明人名称，包括强制性属性 languageCode）、InventorNameLatin（发明人的拉丁名称）、InventionTitle（发明名称）和 SequenceDataQuantity（序列表中序列的总数）。

(3) 序列数据部分的元素

序列数据部分由一个或多个 SequenceData（序列数据）元素组成，每个元素包含一个序列的信息。每个 SequenceData（序列数据）元素必须包含一个强制性属性——sequenceIDNumber（序列识别号）。SequenceData 元素必须包含其依赖元素 INSDSeq（国际核酸序列数据库序列）。

INSDSeq 由 INSDSeq_length（序列长度）、INSDSeq_moltype（分子类型）、INSDSeq_division（序列所属领域，取值“PAT”）、INSDSeq_feature-table（特征表）和 INSDSeq_Sequence（序列）组成。

1）序列长度

序列长度必须公开 INSDSeq_Sequence 元素中包含的序列的核苷酸或氨基酸的数目。

2）分子类型

分子类型必须公开所代表的分子类型。对于核苷酸序列，包括核苷酸类似物序列，分子类型必须表示为 DNA 或 RNA。对于氨基酸序列，分子类型必须表示为 AA。对于含有一个或多个核苷酸的 DNA 和 RNA 区段的核苷酸序列，分子类型必须表示为 DNA。组合的 DNA/RNA 分子必须在特征表中进一步描述，且组合的 DNA/RNA 分子的每个 DNA 和 RNA 区段必须用特征键“misc_feature”和限定词“note”进一步描述，用以指示该区段是 DNA 还是 RNA。

3）特征表

特征表包含关于特定序列内的各个区域的位置和作用信息。除有意跳过的序列外，每个序列都需要一个特征表。INSDSeq_feature-table（特征表）由一个或多个 INSDFeature（特征）元素组成。每个 INSDFeature 元素描述一个特征，由依赖元素 INSDFeature_key（特征关键词）、INSDFeature_location（特征位置）和 INSDFeature_quals（特征限定词）组成。

a. 特征关键词

特征关键词和与其相关限定词共同描述序列的特征。ST.26 标准的特征关键词主要包括核苷酸序列的特征关键词和氨基酸序列的特征关键词，分别

在附件 I 的第 5 和第 7 部分中给出。

b. 特征位置

特征位置必须包含至少一个位置描述符，可以包含一个或多个位置操作符。位置描述符定义了与 INSDSeq_sequence 元素中的序列特征相对应的位点或区域。位置描述符可以是单个残基编号，两个相邻残基编号之间的位点，界定残基编号的连续跨度的区域，或者延伸超出制定残基或残基跨度的位点或区域。当一个特征对应于序列的不联系位点或区域时，多个位置描述符必须使用位置操作符连接。

c. 特征限定词

ST.26 中的特征限定词用于提供除了由特征关键词和特征位置传达的特征之外的信息。限定词有三种类型的值格式（自由文本、受控词表或枚举值、序列）来传达的不同类型的信息。ST.26 标准的特征限定词主要包括核苷酸序列的特征限定词和氨基酸序列的特征限定词，分别在附件 I 的第 6 和第 8 部分给出。

二、 ST.26 标准的发展

(一) 从 CWS 历届会议看 ST.26 的发展

SEQL 工作队成立后围绕着第 44 号任务展开了多轮密集的讨论，取得了实质性的进展，于 2012 年 3 月形成了供各局及其公众进行咨询的标准草案。草案征询公众意见后又进行了讨论与改进，并再次接受审查。SEQL 工作队前后共经过七轮讨论，形成了供 CWS 审议的 ST.26 标准正文及其附件。

2016 年 3 月，CWS 第四届会议续会上，标准委员会正式通过了 ST.26 标准，即“关于用 XML（可扩展标记语言）表示核苷酸和氨基酸序列表的推荐标准”。之后，为了便于落实该标准，SEQL 工作队就如何进一步完善标准及 ST.25 向 ST.26 的过渡事宜进行了多次讨论，并形成修订的 ST.26 标准交由 CWS 第五届会议审议。

2017 年 5 月，CWS 第五届会议上，SEQL 工作队提交了“为 WIPO 标准 ST.25 向 ST.26 的过渡规定制定的建议”的提案；标准委员会通过了经修订的

标准 ST.26（V1.1 版）；并且同意从产权组织标准 ST.25 向 ST.26 的过渡采用“大爆炸”[①]这一方式，以国际申请日作为参考日期，以 2022 年 1 月作为过渡日期。

2018 年 6 月，CWS 第六届会议上，SEQL 工作队提交了关于修订标准 ST.26 的提案，包括对 ST.26 主体部分及其附件一、二、三、四和附件六进行的修订，以及新的附件七（将序列表从 ST.25 转至 ST.26）；标准委员会批准了关于 ST.26 标准的所有修订。

2019 年 7 月，CWS 第七届会议上，SEQL 工作队提出对 ST.26 的附件一至七进行了文字修改，在附件一的核苷酸序列的特征关键词表中新增部分可选限定词；建议将附件三和附件六作为单独文件提供，而标准中仅收入相应的链接。

(二) 从 SEQL 工作队的任务变化看 ST.26 的发展

2010 年 10 月，CWS 第一届会议上设立了第 44 号任务并将该任务分配给了 SEQL 工作队。第 44 号任务最初的描述为“制定一项关于基于可扩展标记语言（XML）的核苷酸和氨基酸序列表表示方法的建议，以作为 WIPO 标准通过。提交这项新 WIPO 标准的提案时，应一并提交报告，说明该标准对现有 WIPO 标准 ST.25 的影响，包括提出对标准 ST.25 的必要修改”。

2016 年 3 月，CWS 第四届会议续会上，标准委员会正式通过了 ST.26 标准，第 44 号任务的描述改为“为 WIPO 标准 ST.25 向 ST.26 的过渡规定制定建议，并在必要时编拟一份关于修订 WIPO 标准 ST.26 的提案。”

2017 年 5 月，CWS 第七届会议上，第 44 号任务的描述改成“为国际局提供支持，提供用户对 ST.26 编著和验证软件工具的要求和反馈意见；在对《PCT 行政规程》进行相应修订的工作上，为国际局提供支持；并且根据标准委员会的要求为产权组织标准 ST.26 编制必要的修订”。

对第 44 号任务的描述的变化能够从一个侧面反映出 ST.26 标准的发展。应标准委员会和各成员国的要求，国际局于 2017 年起开发支持 ST.26 编著和验证的通用软件。截至当前，已基本完成 ST.26 通用工具的开发和验收，

① 所有知识产权局同时从 ST. 25 过渡到 ST. 26。

并进行了最后一轮测试。

(三) 从历史版本看 ST.26 标准的发展

自 2016 年 CWS 第四届续会上，标准委员会通过 ST.26 标准（V1.0 版）以来，ST.26 标准不断地被修订和完善。在 CWS 第五、第六、第七届会议上分别通过了 V1.1、V1.2 和 V1.3 版的修订方案，并于 2019 年 9 月发布了最新版本 V1.3 版。表 2 列出了 ST.26 各版本的修订范围和主要修订内容。从 ST.26 的修订情况可以看出：

- ST.26 标准的修订能够对发现的错误和疏漏及时纠正，同时也注重表述的准确性和清晰度。
- ST.26 标准的修订关注用户的使用，先后新增了附件六（指导文件及其附录）和附件七（关于从 ST.25 转至 ST.26 的建议）。
- ST.26 标准注重与国际标准（ISO 639）、相关产业建议（INSD 等）的一致性。

表 2 ST.26 各版本的修订范围和主要修订内容

版本信息	修订范围	主要修订内容
V1.0	ST.26 主体文件	
	附件一	受控词表。
	附件二	序列表的 DTD。
	附件三	序列表样例（XML 文件）。
	附件四	Unicode 基本拉丁语代码表中的字符子集。
	附件五	附加数据交换需求（仅适用于专利局）。
V1.1	ST.26 主体文件	澄清标准内的肽核酸（PNA）和变体序列；基于欧专局、日本特许厅和美国专商局 2016/2017 举行的公共磋商，完善标准的一般性案文。
	附件一	与 INSD 特征表第 10.6 版保持一致的受控词表。
	附件二	为澄清并与 INSDCDTD 第 1.5 版保持一致，添加评论意见或修改评论意见的措辞。
	附件三	将由两个大写字母组成的语言代码改为小写字母，使其与 ISO 639 中规定的双字母代码相一致，如将“EN”改为“en”。
	新增附件六	指导文件及其附录。
V1.2	ST.26 主体文件	(a)第 7(b)、15、25、27、34 和 95 段中，澄清在序列表中纳入不同序列和进行注释的要求；(b)第 39、43、44 和 46 段中，涉及对 ST.26DTD 的修订；(c)第 55 和 56 段中，更准确地说明核苷酸片段，并澄清相关必要注释；(d)第 81 和 87 段，提高语言的清晰度；以及(e)第 90 段，改正一处错误。

	附件一	(a)第四部分中，删除“和不常见的”和“或不常见的”，因为主体第 3(e)段“修饰氨基酸”中包括“不常见的氨基酸”；(b)第五和第六部分的标题，以“核苷酸”取代“核酸”，以便与 ST.26 的主体保持一致；(c)特征关键词 5.22、5.29、5.31、5.35、5.46、6.55 和 6.56 中，改正当前版本中行文方面疏漏的错误；(d)特征关键词 6.39 和 6.55 中，进行更新以便与最新的 INSDC 特征表更新相对应；以及(e)特征关键词 7.10 中，改正一处疏忽的错误；（f）将三项不同实例中的“合法的”替换为“允许的”。
	附件二	将 INSDFeature_quals 元素使用的非强制性元素 INSDQualifier 改为强制性，即一个 INSDFeature_quals 元素必须具有一个或多个 INSDQualifier 元素（如果有）。
	附件三	根据拟议附件二更新样本，并与 ST.26 的主体保持一致。
	附件四	为清晰起见，更新附件四的标题和介绍性文字。此外，应增补四个疏漏的代码点。
	附件六	更新附录和 XML 序列表；将 15 项不同实例中的“部分”替换为“区域”。
	新增附件七	关于序列表从 ST.25 转至 ST.26 的建议。
V1.3	ST.26 主体文件及附件	对主体文件、附件一至附件七进行部分文字修改。
		更新附件一的表 9，收入 INSDC 特征表 10.8 版中的更新。 附件一的 5.27、5.33 和 5.43 节新增可选限定词。 附件一的 6.16 节的实例和评论部分新增部分内容。
		附件七中部分术语的修改，如将 conversion 改为 transformation，SITE 改为 REGION。
		将附件三和附件六（二者均为 XML 实例）作为单独文件提供，标准中收入相应的链接。

三、 世界主要知识产权局的 ST.26 标准实施计划

(一) 国际局的实施计划

2016 年 3 月，CWS 第四届会议续会上，标准委员会正式通过了 ST.26 标准。2017 年 5 月，CWS 第五届会议上，国际局表示将开发使申请人能够编制和验证符合 ST.26 标准的序列表的通用软件工具，并提出了从 ST.25 向 ST.26 过渡的高级路线图草案。根据国际局的高级路线图草案，ST.26 软件工具的开发将在 2018 年底完成（CWS 第六届会议上，软件开发的完成时间修改为 2019 年）；2019 年 9 月及以后持续的进行软件工具的维护，为知识产权组织及用户提供支持；国际局将在 2018 年至 2020 年间对 PCT 行政规程进行修改；国际局建议各知识产权局 2019 年至 2020 年间对相应的国家法

规进行修改；在 2018 年至 2021 年间对 IT 系统进行升级；2020 年至 2021 年间，国际局将为知识产权组织的员工和申请人提供培训；2022 年 1 月 1 日起 ST.26 标准全面实施。

(二) 美国的实施计划

美国专利商标局 USPTO 计划于 2017 年至 2021 年进行内部系统置换；2019 年整合测试 WIPO 的验证工具；2019 年至 2021 年修改国内条款及内部员工信息技术培训；2020 年至 2022 年进行申请人的信息技术培训；并将于 2022 年开始全面实施 ST.26 标准。

(三) 欧洲的实施计划

欧洲专利局 EPO 于 2018 年为 WIPO 研发 ST.26 工具提供支持，2019 年从 IT 技术层面分析 ST.25 向 ST.26 转换的影响，2019 年下半年至 2020 年修改工作流程和工具以适应 ST.26，2020 年底至 2021 年对欧洲专利法规相关条款进行修改，2021 年引导申请人试用 ST.26，并将于 2022 年接收符合 ST.26 标准的申请。

(四) 日本的实施计划

日本专利特许厅 JPO 的实施计划为：2017 年至 2021 年底进行内部系统的全面改革，包括审查系统和出版系统，2019 年为计划阶段；2020 年至 2021 年依据 ST.26 软件为序列表调整申请系统，必要的话会修改国内条款，并帮助申请人做好符合 ST.26 标准申请的准备；并将于 2022 年接收符合 ST.26 标准的申请。

(五) 韩国的实施计划

韩国工业产权局 KIPO 计划于 2017 年调查用户对于 WIPO ST.26 的软件需求；2018 年进行软件研发，修改申请系统和审查系统为 ST.26 申请做准备；2019 年启动 WIPO ST.26 软件分配，为用户提供软件测试；2019 年至 2021 年持续的进行系统（申请系统、审查系统等）改进；2020 年至 2022

年推进标准从 ST.25 向 ST.26 过渡，并为就如何使用 WIPO 的 ST.26 软件为用户提供培训；2021 年下半年开始为感兴趣的国内申请人提供 ST.26 标准申请的测试，并开始接收符合 ST.26 标准的序列表；并将于 2022 年全面接收符合 ST.26 标准的申请。

(六) 中国的实施建议

根据国际局提供的 ST.25 向 ST.26 过渡的路线图，借鉴世界主要知识产权局的实施计划，结合中国的具体实际，建议中国在 2022 年 1 月 1 日来临之前做好以下工作：

- 制定符合中国现实需求的 ST.26 标准的实施计划。
- 积极参与 ST.26 标准软件工具的测试，认真研判 ST.26 软件工具与申请、审查、检索系统的对接，在智能化系统升级过程中充分考虑 ST.26 标准及其配套软件的应用。
- 根据 PCT 行政规程的修改，对国内和国际申请的规则和行政规程进行修订。
- 为申请人（含代理机构）、审查员及其他用户提供关于 ST.26 标准及其配套工具，以及它们在中国的申请、审查和检索系统中的实现方式的培训。
- 为申请人提供中文版 ST.26 工具的使用手册。

第三章 ST.26 标准与 ST.25 标准的比较研究

本章在梳理 ST.25 标准的基础上，对 ST.26 标准和 ST.25 标准从适用范围、标准格式、标准内容等方面进行全面比较和系统分析，并讨论标准变化对智能化升级系统设计的影响。

一、 ST.25 标准简介

1998 年 11 月，WIPO 出台了《专利申请中核苷酸和氨基酸序列表的表述标准》，即 ST.25 标准。该标准是适用于专利合作条约（PCT）的国际申请中核苷酸与氨基酸序列表的表述标准，其附属于 PCT 行政规程的附件 C。世界各国专利局需要对该标准作必要修改后方可适用于非 PCT 国际申请中的所有专利申请。

2009 年 10 月，在 WIPO 信息与技术标准委员会（SCIT）标准与文献工作组（SDWG）的第 11 次会议上，讨论并通过了 ST.25 标准的修订意见(即 33 号任务)，删除了“混合模式序列表申请”；以及要求含有序列表有关表格的页在确定申请费时计算在内，不论其是否以电子方式提交；并对以 ST.25 文本格式提交的序列表免除收费。

(一) ST.25 标准的目标

ST.25 标准是为了使国际专利申请中核苷酸和氨基酸序列表的表述标准化而制定的，其目标是使申请人制作的序列表能在国际阶段被所有受理局、国际检索和初步审查单位以及在国家阶段被所有指定局和选定局接受。

ST.25 标准的目的在于：

- 提高国际申请中提供的核苷酸和氨基酸序列的表述精确度和质量；
- 使申请人、公众和审查员更容易表述和传播该序列；
- 有利于序列数据的检索；
- 允许电子形式的序列数据的交换以及将序列数据输入计算机数据库。

(二) ST.25 标准的结构

ST.25 标准由主体文件和三个附件组成。ST.25 标准中的主体文件主要对序列表、核苷酸序列及氨基酸序列使用的符号和格式、序列表中其他有用的信息（数据元素、特征的表述、自由文本等）、自由文本在说明书主体部分的重复、序列表的电子形式等进行了规范。附件 1 为数字标识符；附件 2 为核苷酸和氨基酸符号和特征表，主要包括核苷酸表、修饰的核苷酸表、氨基酸表、修饰的氨基酸和非常见氨基酸表、与核苷酸序列相关的特征关键词表、与蛋白质序列相关的特征关键词表；附件 3 为序列表样例。

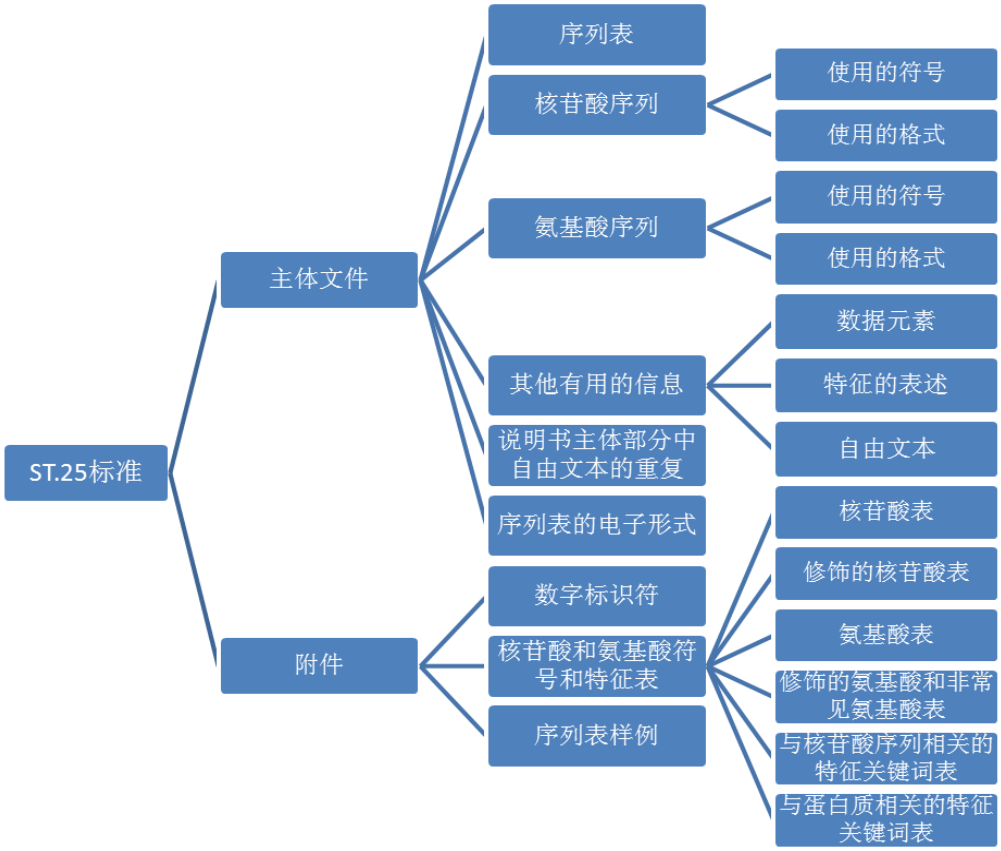


图 3 ST.25 标准内容的结构示意图

(三) ST.25 标准的序列

ST.25 标准主要从序列的表示、序列表的元素和特征的表示几方面对专利申请中核苷酸和氨基酸序列表的表示给出了建议。符合 ST.25 标准的序列表中的序列不包括具有少于 10 个具体定义的核苷酸或少于 4 个具体定义的氨基酸的序列。

1. 序列的表示

对于核苷酸序列和氨基酸序列，ST.25 分别从使用的符号和使用的格式两个方面进行了说明。

2. 序列表的元素

ST.25 中用数字标识符对各数据项进行标记。表 3 列出了 ST.25 中的各项元素信息及其对应的数字标识符。从表 3 可以看出，ST.25 中涉及的强制性元素主要包括申请信息：<110>申请人、<120>发明名称、<130>案卷参考号、<140>专利申请号、<141>专利申请日，优先权信息：<150>优先权号、<151>优先权日、序列信息：<160>序列个数、<210>序列标识符、<211>序列长度、<212>序列类型、<213>生物体、<220>特征、<221>名称/关键词、<222>位置、<223>其他信息和<400>序列。

特征（<220>）使用附件 2 中表 5（核苷酸序列相关的特征关键词表）和表 6（蛋白质序列相关的特征关键词表）的“特征关键词”来描述。当序列中存在“n”或“Xaa”或修饰的碱基或修饰的/异常的 L-氨基酸时，特征部分必须用数据元素——关键词（<221>）、位置（<222>）和其他信息（<223>）进行限定；当生物体（<213>）为“人工序列”或“未知”时，必须用数据元素——特征（<220>）和其他信息（<223>）进行限定。<223>是用“自由文本”描述序列特征，自由文本优选地使用英语撰写，对于给定的数据元，自由文本的长度不超过 4 行，每行最多 65 个字符。

表 3 ST.25 标准中元素信息统计表

数字标识符	字段名称	是否强制性元素
<110>	申请人名称	M
<120>	发明名称	M
<130>	案卷参考号	M
<140>	专利申请号	M
<141>	专利申请日	M
<150>	优先权号	M
<151>	优先权日	M
<160>	序列个数	M
<170>	软件版本	O
<210>	序列标识符	M

<211>	序列长度	M
<212>	序列类型	M
<213>	生物体	M
<220>	特征	M
<221>	名称/关键词	M
<222>	位置	M
<223>	其它信息	M
<300>	公开出版信息	O
<301>	作者	O
<302>	题目	O
<303>	杂志名称	O
<304>	卷号	O
<305>	出版号	O
<306>	页码	O
<307>	出版日期	O
<308>	数据库登记号	O
<309>	录入数据库日期	O
<310>	专利公开号	O
<311>	专利申请日	O
<312>	专利公开日	O
<313>	相关残基	O
<400>	序列	M
注：M 表示强制性元素，O 表示可选元素。		

二、 ST.26 标准与 ST.25 标准的比较

(一) 标准适用范围的比较

ST.25 标准主要适用于 PCT 国际专利申请，若要用于非 PCT 国际申请的专利申请，则需要对该标准进行修改后才可能适用[①]。

ST.26 标准适用于国际、国内或区域程序的专利申请。

(二) 标准格式的比较

ST.25 标准为 TXT 格式，而 ST.26 标准采用了 XML 格式，在数据管理、交换和解析方面更加安全、便捷。

① 对 PCT 特定的某些条款和要求可能不适用于非 PCT 国际申请的专利申请，如果各知识产权局的国内法和实践中采用的做法与 PCT 行政规程不一致，则可以不遵循此标准。

(三) 标准内容的比较

1. 标准元素的比较

ST.26 包括通用数据部分和序列数据部分。对通用数据部分而言，ST.26 包含了 ST.25 中的所有数据项，且比 ST.25 新增了“序列表文件的产生日期”、“发明人”及其强制性属性等信息。对序列数据部分而言，ST.26 新增了“限定词”相关的信息。此外，缺少 ST.25 中“公开/出版信息”相关的数据项。

ST.25 中用数字标识符对各种类型的数据进行标记，ST.26 中使用英语中的术语作为元素名称和属性名称，对数据进行标记。

表 4 列出了 ST.25 和 ST.26 的元素对照表。从表 4 可以看出：

- 关于申请人名称，ST.25 中要求非拉丁字母的申请人名要音译或英译成拉丁字母表示；ST.26 中则在申请人名称元素下增加了“ApplocantNameLatin”字段，当申请人名称中包含非拉丁字母时，该字段为强制性的。
- 对于知识产权局代码，ST.25 中是在专利申请号前标注知识产权局代码，ST.26 中在申请标识信息元素下增加了“IPOfficeCode”元素。
- 对于优先权信息，ST.25 中给出的是在先申请的申请号和申请日信息，ST.26 中给出的则是最早优先权信息。
- 关于序列的分子类型，ST.25 中取值为 DNA，RNA 或 PRT；ST.26 中取值为 DNA，RNA 或 AA。在核苷酸序列包含 DNA 和 RNA 片段的情况下，ST.25 要求<212>中的值应为“DNA”，并在<220>至<223>特征部分进一步描述组合的 DNA/RNA 分子；然而，进一步描述的确切性不明确；ST.26 要求组合 DNA/RNA 分子的每个 DNA 和 RNA 区段都必须使用特征关键词“misc_feature”（包括区段的位置）和限定词“note”（表明区段是 DNA 还是 RNA）来进一步描述。
- 对于“生物体”的表述，ST.25 中给出了“生物体”字段，而 ST.26 中“生物体”是通过特征关键词及限定词来表述的。在核苷酸列表中用特征关键词 source 及其强制性限定词 organism 和 mol_type 表述，在氨基酸列表中，用特征关键词 SOURCE 及其强制性限定词 ORGANISM 和 MOL_TYPE 表述。

表 4 ST.25 和 ST.26 的元素对照表

信息项	元素/属性	ST.25		ST.26		备注
		有无	是否强制	有无	是否强制	
序列表信息	DTD 版本	×	‘--	√	M	生成序列表文件的 DTD 版本
	软件版本	√	O	√	O	生成序列表文件的软件版本
	软件名称	×	‘--	√	O	生成序列表文件的软件名称
	文件名称	×	‘--	√	O	序列表文件的名称
	生产日期	×	‘--	√	O	序列表文件产生的日期
申请信息	申请人名称	√	M	√	M	ST.25 中非拉丁字母的申请人名要译成拉丁字母表示；ST.26 中分配申请号后的任何时间提供序列表时，为强制性元素。
	申请人拉丁名称	×	‘--	√	M	申请人名包括非拉丁字母时，为强制性元素。
	发明人名称	×	‘--	√	O	
	语言代码	×	‘--	√	M	属性
	发明人拉丁名称	×	‘--	√	O	
	发明名称	√	M	√	M	
	案卷参考号	√	M/O	√	M/O	当序列表在分配申请号之前的任一时间提供时为强制性元素，其余情况为可选元素。
	专利申请号	√	M	√	M	
	知识产权局代码	√	M	√	M	ST.25 中在申请号前标注知识产权局代码
	专利申请日	√	M	√	M	分配申请号后的任何时间提供序列表时，为强制性元素
优先权信息	优先权号	√	M	√	M	要求优先权时，为强制性元素。
	优先权国别代码	×	‘--	√	M	要求优先权时，为强制性元素。
	优先权日	√	M	√	M	要求优先权时，为强制性元素。

序列信息	序列	√	M	×	M	
	序列特征表	×	‘--	√	M	序列的注释列表
	序列表中序列的个数	√	M	√	M	
	序列标识符	√	M	√	M	
	序列的长度	√	M	√	M	序列长度用碱基对或氨基酸的数量表示
	序列的分子类型	√	M	√	M	ST.25 中取值为 DNA,RNA 或 PRT；ST.26 中取值为 DNA，RNA 或 AA
	序列分属领域	×	‘--	√	M	ST.26 中为强制性元素，取值为“PAT”
特征信息	生物体	√	M	√	M	ST.26 中对于生物体通过特征关键词及限定词来表述。在核苷酸列表中用特征关键词 source 及其强制性限定词 organism 和 mol_type 表述，在氨基酸列表中，用特征关键词 SOURCE 及其强制性限定词 ORGANISM 和 MOL_TYPE 表述。
	特征	√	M	√	M	ST.25 中的序列使用了 n 或 Xaa 或修饰的碱基或修饰的/异常 L-氨基酸，或者生物体是“人工序列”或者“未知”时，为强制性元素
	名称/关键词	√	M	√	M	ST.25 中的序列使用了 n 或 Xaa 或修饰的碱基或修饰的/异常 L-氨基酸时，为强制性元素
	位置	√	M	√	M	
	INSD 特征_限定词	×	‘--	√	M	当特征关键词需要限定词时，为强制性元素，其他情形为非强制性元素。
	INSD 特征_限定词名称	×	‘--	√	M	
	INSD 特征_限定词取值	×	‘--	√	M	
	其它信息	√	M	√	O	ST.25 中的序列使用了 n 或 Xaa 或修饰的碱基或修饰的/异常 L-氨基酸，或者生物体是“人工序列”或者“未知”时，为强制性元素
公开/出版信息	作者	√	O	×	‘--	

	题目	√	O	×	‘--	
	杂志名称	√	O	×	‘--	
	公开出版物的卷号	√	O	×	‘--	
	公开出版物的出版号	√	O	×	‘--	
	页码	√	O	×	‘--	
	出版日期	√	O	×	‘--	
	公开出版物的数据库登记号	√	O	×	‘--	
	录入数据库的日期	√	O	×	‘--	
	专利公开号	√	O	×	‘--	
	专利申请日	√	O	×	‘--	
	专利公开日	√	O	×	‘--	
	相关残基	√	O	×	‘--	
注：（1）元素/属性列，左对齐为“元素”，右对齐为“属性”；（2）M 表示强制性元素/属性，O 表示可选元素/属性；（3）‘-- 表示标准中没有相应元素/属性是否为强制的规定。						

2. 序列的比较

(1) 序列的范围

ST.25 和 ST.26 中均对核苷酸和氨基酸序列长度做出了限制，即不包括少于 10 个特别定义的核苷酸或 4 个特别定义的氨基酸的序列。但是，ST.26 明确要求包括:（a）分支序列；(b)具有 D-氨基酸的序列；（c）核苷酸类似物；和(d)具有无碱基位点的序列。在 ST.25 中，明确不包括（a）分支序列，而对包括或禁止（b）-（d）序列的要求并不明确。此外，在 ST.26 中，肽核酸（PNA）被认为是核苷酸而不是氨基酸，而 ST.25 中对肽核酸未做明确说明。

(2) 序列的表示

ST.25 和 ST.26 中都要求，每个序列必须被分配独立的序列标识号。序列标识号从 1 开始，按照整数连续递增。当序列标识号没有对应的序列出现时（如“有意跳过的序列”），序列位置应用“000”。序列的总数必须在序列表中指出，且必须等于序列标识号的总数。

1）核苷酸序列的表示

核苷酸序列必须按照 5'端至 3'端方向从左至右的单链表示，5'和 3'或任何其它类似的指定不得包括在序列中。

序列中的核苷酸必须使用核苷酸序列表中的单个小写字母代码表示。

a. 双链核苷酸序列的表示

对于双链核苷酸序列，ST.25 中未做出明确要求；ST.26 中规定：(a)当双链完全互补时，表示为单个序列或两个单独的序列，并且每个序列分配其自己的序列标识号；(b)当双链不完全互补时，表示为两个单独的序列，且每个序列分配其自己的序列标识号。

b. 修饰的核苷酸序列的表示

ST.25 中，如果修饰的核苷酸是修饰的核苷酸列表中的核苷酸之一，则序列中修饰的核苷酸应表示为相应的未修饰的核苷酸或者“n”，并在序列表的特征部分进一步描述。

ST.26 中，在可能的情况下，修饰的核苷酸应当用相应的未修饰的核苷

酸进行表示，并且应当在特征表中使用特征关键词“modified_base”和强制性限定词“mod_base”结合受控词表中的限定词取值做出进一步描述。

ST.26 中，DNA 中的“尿嘧啶”或 RNA 中的“胸腺嘧啶”被认为是修饰的核苷酸，序列中必须用“t”表示，并在特征表中使用特征关键词“modified_base”进一步描述，并分别用限定词“mod_base”并以“OTHER”作为限定词的取值，和限定词“note”并以“尿嘧啶”或“胸腺嘧啶”作为限定词的取值。

c.未知的核苷酸序列的表示

ST.25 和 ST.26 中未知的核苷酸均用“n”表示，且 ST.26 中需要在特征表中使用特征关键词“unsure”做出进一步描述。

2）氨基酸序列的表示

氨基酸序列中的氨基酸必须从左到右以氨基到羧基的方向列出，氨基和羧基基团不得在序列中表示。

ST.25 中氨基酸应当使用第一个字母大写的三字母代码表示。ST.26 中氨基酸必须用氨基酸序列表单个大写字母表示。

a. 含有空白或内部终止符号的氨基酸序列的表示

ST.25 中，含有空白或内部终止符号（“ter”或“*”或“.”）的氨基酸序列可以不表示为单个氨基酸序列，但是应该作为单独的氨基酸序列列出。

ST.26 中，由空白或内部终止符号（“ter”或“*”或“.”）分开的氨基酸序列必须作为氨基酸序列的单独序列包括在内，且每个被分开的序列必须被分配其自己的序列标识号，且终止子符号和空白不得包括在序列表中的序列中。

b. 修饰的氨基酸序列的表示

ST.25 中，如果修饰的氨基酸是修饰的氨基酸列表中的氨基酸之一，则序列中修饰的氨基酸和异常的氨基酸应表示为相应的未修饰的氨基酸或者“Xaa”，并在序列表的特征部分进一步描述。

ST.26 中，在可能的情况下，修饰的氨基酸应当用相应的未修饰的氨基酸或者“X”进行表示，并且必须在特征表中进一步描述。

c.未知的氨基酸序列的表示

ST.25 中“Xaa”表示未知的或其他的氨基酸。ST.26 中，未知的氨基酸在序列中用“X”表示，且在特征表中使用特征关键词“UNSURE”和可选限定词“NOTE”做出进一步描述。

3）特殊序列的表示

a. 变体序列的表示

关于变体序列，ST.25 中未做出明确规定。ST.26 中，通过枚举残基的方式公开的基础序列及其任何变体必须各自包括在序列表中并分配它们自己的序列标识号；以在一个或多个位置带有枚举的替代的变体残基的单个序列公开的变体序列必须包括在序列表中，并以单个序列标识；仅通过序列表中基础序列的删除、插入或取代而公开的变体序列应当包括在序列表中。特定变体的序列的注释必须包括特征键、限定词，以及特征位置。

b. 编码序列的表示

由编码序列编码的氨基酸序列和限定词“translation”中公开的序列必须包括在序列表中，并分配其自身的序列标识符，分配的序列标识符必须作为具有“CDS”特征关键词的限定词“protein_id”的取值中提供。

c. 人工源序列的表示

对于人工源序列，ST.25 的<213>字段“Organism”用“Artificial Sequence”描述，ST.26 中用特征关键词“source”或“SOURCE”及其特征限定词“organism”或“ORGANISM”进行描述，且“organism”或“ORGANISM”的取值为“synthetic construct”。

d. 未知源序列的表示

对于生物体源的科学名称未知的序列，ST.25 的<213>字段“Organism”用“Unknown”描述，ST.26 中用特征关键词“source”或“SOURCE”及其特征限定词“organism”或“ORGANISM”进行描述，且“organism”或“ORGANISM”的取值为“unidentified”。

e. 片断组成的序列的表示

ST.25 中规定由较大序列的一个或多个非连续的片断或者不同序列的片

断组成的氨基酸/核苷酸序列应该作为带有单独序列标识符的一个单独序列编号。

ST.26 中规定由较大序列的一个或多个非连续的片断或者不同序列的片断必须包含在序列表中并分配其自己的序列标识号。

f. 缺口分开的序列的表示

ST.25 中规定具有一个或多个缺口的序列应作为具有不同序列标识符的多个单独序列进行编号，而单独序列的数目在数值上等于序列数据中连续链的数目。

ST.26 中对含有由未知或未公开数目的残基的一个或多个缺口分隔的特别定义的残基的区域的序列不得在序列表中表示为单个序列。

g. “n”或“X”残基分开的序列的表示

ST.26 中，如果一个序列含有由连续“n”或“X”残基的一个或多个区域分开的特别定义的残基的区域，且其中每个区域中“n”或“X”残基的确切数目已被公开，则该序列必须作为一个序列包括在序列表中并分配其自身的序列标识号。

(3) 序列的编号

1）核苷酸序列的编号

核苷酸的编号应该从序列中呈现的第一个核苷酸开始编号为 1，从 5’到 3’的方向上贯穿整个序列连续编号，且最后一个残基位置的编号必须等于序列中核苷酸的数目。

对于环状构型的核苷酸序列，ST.25 中的第一个核苷酸可以由申请人指定，其余编号规则与普通核苷酸类似；ST.26 中申请人必须选择残基位置编号 1 的核苷酸。

2）氨基酸序列的编号

ST.25 标准允许氨基酸的编号可以包括负数，而 ST.26 标准不允许特征位置有负数。

对于环状构型的氨基酸序列，ST.25 中第一个氨基酸可以由申请人指定，其余编号规则与普通氨基酸类似；ST.26 中申请人必须选择残基位置编号为

1 的氨基酸，之后按照从氨基到羧基的方向连续编号。

(4) 序列的代码

在 ST.25 和 ST.26 两个标准中，序列的代码通过核苷酸序列、修饰的核苷酸序列、氨基酸序列和修饰的氨基酸序列的受控词表加以呈现。

1）核苷酸序列的受控词表

表 5 给出了 ST.25 和 ST.26 核苷酸序列的受控词对照表。从表 5 可以看出：

- ST.25 标准包含 16 个核苷酸符号，ST.26 标准包含 15 个核苷酸符号。主要区别在于 ST.26 标准比 ST.25 标准中少一个符号“u”。
- ST.25 中，用“t”表示 DNA 序列中出现的“胸腺嘧啶（thymine）”，“u”表示 RNA 序列中出现的“尿嘧啶（uracil）”。ST.26 中，DNA 序列中的“胸腺嘧啶（thymine）”和 RNA 序列中的“尿嘧啶（uracil）”均用“t”表示。

表 5 ST.25 和 ST.26 核苷酸序列的受控词对照表

序号	核苷酸符号	含义	ST.25	ST.26	备注
1	a	adenine	√	√	
2	g	guanine	√	√	
3	c	cytosine	√	√	
4	t	thymine	√	√	在 ST.26 标准中，若无进一步说明，则“t”代表 DNA 中的胸腺嘧啶和 RNA 中的尿嘧啶
5	u	uracil	√	×	
6	r	a or g	√	√	
7	y	c or t/u	√	√	
8	m	a or c	√	√	
9	k	g or t/u	√	√	
10	s	g or c	√	√	
11	w	a or t/u	√	√	
12	b	c or g or t/u; not a	√	√	
13	d	a or g or t/u; not c	√	√	
14	h	a or c or t/u; not g	√	√	
15	v	a or c or g; not t/u	√	√	

16	n	a or c or g or t/u; “unknown” or “other”	√	√	

2）修饰的核苷酸序列的受控词表

表 6 给出了 ST.25 和 ST.26 修饰的核苷酸序列的受控词对照表。从表 6 可以看出：

- ST.25 中修饰的核苷酸符号为 46 个， ST.26 中修饰的核苷酸符号为 48 个。二者的主要区别在于：(a). 二氢尿苷（dihydrouridine）在 ST.25 中用符号“d”表示，在 ST.26 中用“dhu”表示；(b). 5-甲基尿苷（5-methyluridine）在 ST.25 中用符号“t”表示，在 ST.26 中用“m5u”表示；(c).ST.26 中增加了“m4c”和“OTHER”两个符号，其中“m4c”表示 N4-甲基胞嘧啶（N4-methylcytosine），“OTHER”用于表示受控词表中未出现的修饰的核苷酸，且要与限定词“note”一起使用。
- ST.26 中修饰的核苷酸序列的符号仅用作限定词“mod_base”的取值。如果修饰的核苷酸未出现在受控词表中，则必须用“OTHER”作为“mod_base”的取值，且在限定词“note”中提供修饰的核苷酸的完整的、未缩写的全名。

表 6 ST.25 和 ST.26 修饰的核苷酸序列的受控词对照表

序号	修饰核苷酸符号	ST.25	ST.26	备注
1	ac4c	√	√	
2	chm5u	√	√	
3	cm	√	√	
4	cmnm5s2u	√	√	
5	cmnm5u	√	√	
6	d	√	×	二氢尿苷(dihydrouridine)在 ST.26 中用 dhu 表示，在 ST.25 中用 d 表示
7	fm	√	√	
8	gal q	√	√	
9	gm	√	√	
10	i	√	√	
11	i6a	√	√	
12	m1a	√	√	

13	m1f	√	√	
14	m1g	√	√	
15	m1i	√	√	
16	m22g	√	√	
17	m2a	√	√	
18	m2g	√	√	
19	m3c	√	√	
20	m5c	√	√	
21	m6a	√	√	
22	m7g	√	√	
23	mam5u	√	√	
24	mam5s2u	√	√	
25	man q	√	√	
26	mcm5s2u	√	√	
27	mcm5u	√	√	
28	mo5u	√	√	
29	ms2i6a	√	√	
30	ms2t6a	√	√	
31	mt6a	√	√	
32	mv	√	√	
33	o5u	√	√	o5u 在 ST.26 中表示 uridine-5-oxyacetic acid (v)，在 ST.25 中表示 uridine-5-oxyacetic acid
34	osyw	√	√	
35	p	√	√	
36	q	√	√	
37	s2c	√	√	
38	s2t	√	√	
39	s2u	√	√	
40	s4u	√	√	
41	t	√	×	5-甲基尿苷 (5-methyluridine)在 ST.26 中用 m5u 表示，在 ST.25 中用 t 表示
42	t6a	√	√	
43	tm	√	√	
44	um	√	√	
45	yw	√	√	
46	x	√	√	
47	OTHER	×	√	需要注释限定符

48	m4c	×	√	N4-methylcytosine
49	dhu	×	√	dihydrouridine
50	m5u	×	√	5-methyluridine

3）氨基酸序列的受控词表

表 7 给出了 ST.25 和 ST.26 标准中氨基酸序列的受控词对照表。从表 7 可以看出：

- ST.25 标准中，氨基酸用首字母大写的三字母代码表示；ST.26 标准中，所有氨基酸必须用单个大写字母表示，且用于表示氨基酸的任何符号是仅一个残基的等同物。
- ST.25 标准包括 23 个氨基酸符号，ST.26 标准包括 26 个氨基酸符号。与 ST.25 标准相比，ST.26 标准增加了 3 个氨基酸符号：“O”-表示吡咯赖氨酸（Pyrrolysine）、“U”-表示硒半胱氨酸（Selenocysteine）、“J”-表示亮氨酸（Leucine）或异亮氨酸（Isoleucine）。
- ST.25 标准中不提供 Xaa 的默认值，若一个序列包括 Xaa，ST.25 要求字段<223>中包括关于该残基的进一步描述，以及字段<221>（特征名称）和<222>（特征位置）；ST.26 标准中提供了 X 的默认值，并不总是要求提供进一步信息。

表 7 ST.25 和 ST.26 氨基酸序列的受控词对照表

序号	ST.25		ST.26	
	符号	氨基酸	符号	氨基酸
1	Ala	Alanine	A	Alanine
2	Cys	Cysteine	C	Cysteine
3	Asp	Aspartic Acid	D	Aspartic acid (Aspartate)
4	Glu	Glutamic Acid	E	Glutamic acid (Glutamate)
5	Phe	Phenylalanine	F	Phenylalanine
6	Gly	Glycine	G	Glycine
7	His	Histidine	H	Histidine
8	Ile	Isoleucine	I	Isoleucine
9	Lys	Lysine	K	Lysine
10	Leu	Leucine	L	Leucine
11	Met	Methionine	M	Methionine
12	Asn	Asparagine	N	Asparagine
13	Pro	Proline	P	Proline
14	Gln	Glutamine	Q	Glutamine
15	Arg	Arginine	R	Arginine
16	Ser	Serine	S	Serine

17	Thr	Threonine	T	Threonine
18	Val	Valine	V	Valine
19	Trp	Tryptophan	W	Tryptophan
20	Tyr	Tyrosine	Y	Tyrosine
21	Asx	Asp or Asn	B	Aspartic acid or Asparagine
22	Glx	Glu or Gln	Z	Glutamine or Glutamic acid
23	Xaa	unknown or other	U	Selenocysteine
24			O	Pyrrolysine
25			J	Leucine or Isoleucine
26			X	A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V; “unknown” or “other

4）修饰的氨基酸序列的受控词表

表 8 给出了 ST.25 和 ST.26 修饰的氨基酸序列的受控词对照表。从表 8 可以看出：

两个标准中，修饰的氨基酸符号没有区别。在 ST.26 中，修饰的氨基酸受控词表中的符号仅用于特征关键词“MOD_RES”和“SITE”的强制性限定词“NOTE”的取值。当修饰的氨基酸在受控词表中未出现时，应使用完整的、未缩写的修饰的氨基酸全称作为强制性限定词“NOTE”的取值。ST.25 标准对受控词表中未出现的修饰的氨基酸，则不需要使用完整的、未缩写的名称。

表 8 ST.25 和 ST.26 修饰的氨基酸序列的受控词对照表

序号	ST.25		ST.26	
	符号	修饰的氨基酸	符号	修饰的氨基酸
1	Aad	2-Aminoadipic acid	Aad	2-Aminoadipic acid
2	bAad	3-Aminoadipic acid	bAad	3-Aminoadipic acid
3	bAla	beta-Alanine, beta-Aminopropionic acid	bAla	beta-Alanine, beta-Aminoproprionic acid
4	Abu	2-Aminobutyric acid	Abu	2-Aminobutyric acid
5	4Abu	4-Aminobutyric acid, piperidinic acid	4Abu	4-Aminobutyric acid, piperidinic acid
6	Acp	6-Aminocaproic acid	Acp	6-Aminocaproic acid
7	Ahe	2-Aminoheptanoic acid	Ahe	2-Aminoheptanoic acid
8	Aib	2-Aminoisobutyric acid	Aib	2-Aminoisobutyric acid
9	bAib	3-Aminoisobutyric acid	bAib	3-Aminoisobutyric acid
10	Apm	2-Aminopimelic acid	Apm	2-Aminopimelic acid
11	Dbu	2,4 Diaminobutyric acid	Dbu	2,4-Diaminobutyric acid

12	Des	Desmosine	Des	Desmosine
13	Dpm	2,2’-Diaminopimelic acid	Dpm	2,2’-Diaminopimelic acid
14	Dpr	2,3-Diaminopropionic acid	Dpr	2,3-Diaminoproprionic acid
15	EtGly	N-Ethylglycine	EtGly	N-Ethylglycine
16	EtAsn	N-Ethylasparagine	EtAsn	N-Ethylasparagine
17	Hyl	Hydroxylysine	Hyl	Hydroxylysine
18	aHyl	allo-Hydroxylysine	aHyl	allo-Hydroxylysine
19	3Hyp	3-Hydroxyproline	3Hyp	3-Hydroxyproline
20	4Hyp	4-Hydroxyproline	4Hyp	4-Hydroxyproline
21	Ide	Isodesmosine	Ide	Isodesmosine
22	aIle	allo-Isoleucine	aIle	allo-Isoleucine
23	MeGly	N-Methylglycine, sarcosine	MeGly	N-Methylglycine, sarcosine
24	MeIle	N-Methylisoleucine	MeIle	N-Methylisoleucine
25	MeLys	6-N-Methyllysine	MeLys	6-N-Methyllysine
26	MeVal	N-Methylvaline	MeVal	N-Methylvaline
27	Nva	Norvaline	Nva	Norvaline
28	Nle	Norleucine	Nle	Norleucine
29	Orn	Ornithine	Orn	Ornithine

(5) 序列中歧义符号的使用

ST.25 中未对歧义符号做出明确规定。ST.26 标准中规定：如果需要使用歧义符号（代表可替代的两个或多个氨基酸/核苷酸），则应使用最严格的符号。例如，在氨基酸序列中，如果给定位置的氨基酸可以是天冬氨酸或天冬酰胺，则应使用符号“B”，而不是“X”。因为除进一步说明外，X 可以被解释为“A”，“R”，“N”，“D”，“C”，“Q”，“E”，“G”，“H”，“I”，“L”，“K”，“M”，“F”，“P”，“O”，“S”，“U”，“T”，“W”，“Y”或“V”中的任何一个。在核苷酸序列中，如果给定位置的碱基可以是“a 或 g”的情况下，则应使用“r”，而不是“n”，因为符号“n”在没有进一步描述的情况下使用时将被解释为“a 或 c 或 g 或 t / u”。

3. 特征关键词的比较

ST.25 中每个特征关键词具有<222>字段以表明特征位置，但是对大多数特征来说，ST.25 不要求位置的指示，且位置信息的格式不是标准化的，也没有位置操作符。ST.26 中具有标准化的位置描述符和操作符，每个特征必须包含至少一个位置描述符。

(1) 核苷酸序列的特征关键词表的比较

表 9 列出了 ST.25 和 ST.26 核苷酸序列的特征关键词对照表，从表 9 可

以看出：

ST.25 中有，ST.26 中没有的核苷酸序列特征关键词共 23 个，主要区别是 ST.25 的特征关键词在 ST.26 中为某个特征关键词限定词或者限定词的取值。ST.25 中没有，ST.26 中有的核苷酸序列特征关键词共 9 个，分别是 centromere（着丝粒）、mobile_element（移动元素）、ncRNA（非编码 RNA）、operon（操纵子）、oriT（转移原点）、propeptide（前肽）、regulatory（调节）、telomere（端粒）和 tmRNA（转录信使 RNA）。

ST.25 序列特征关键词表包括“关键词”和对关键词的“说明”；ST.26 序列表特征关键词表中包括关键词、关键词定义、可选限定词、(强制性限定词词、生物体范围、分子范围、补充说明等）内容，其中关键词、关键词定义、可选限定词为所有关键词都出现的项，其他项则不是所有关键词都有相应的信息。

表 9 ST.25 和 ST.26 核苷酸序列的特征关键词对照表

序号	核苷酸序列的特征关键词	ST.25	ST.26	备注
1	C_region	√	√	
2	CDS	√	√	
3	centromere	×	√	被鉴定为着丝粒并且已经通过实验表征的生物感兴趣区域。
4	D_loop	√	√	
5	D_segment	√	√	
6	exon	√	√	
7	gene	√	√	
8	iDNA	√	√	
9	intron	√	√	
10	J_segment	√	√	
11	mat_peptide	√	√	
12	misc_binding	√	√	
13	misc_difference	√	√	
14	misc_feature	√	√	
15	misc_recomb	√	√	
16	misc_RNA	√	√	
17	misc_structure	√	√	
18	mobile_element	×	√	含有可移动元件的基因组区域。
19	modified_base	√	√	
20	mRNA	√	√	
21	ncRNA	×	√	非蛋白编码基因，除了核糖体 RNA 和转移 RNA，其功能分子为 RNA 转录物
22	N_region	√	√	

23	operon	×	√	包含多顺反子转录物的区域，所述多顺反子转录物包括在相同调节序列/启动子控制下且在相同生物途径中的基因簇。
24	oriT	×	√	转录起源；在结合或移动过程中启动转录的 DNA 分子区域。
25	polyA_site	√	√	
26	precursor_RNA	√	√	
27	prim_transcript	√	√	
28	primer_bind	√	√	
29	propeptide	×	√	前肽编码序列；前蛋白的结构域的编码序列，其被切割以形成成熟蛋白产物。
30	protein_bind	√	√	
31	regulatory	×	√	在转录、翻译、复制或染色质结构的调节中起作用的序列的任何区域。
32	repeat_region	√	√	
33	rep_origin	√	√	
34	rRNA	√	√	
35	S_region	√	√	
36	sig_peptide	√	√	
37	source	√	√	
38	stem_loop	√	√	
39	STS	√	√	
40	telomere	×	√	鉴定为端粒并且已经通过实验表征的生物学感兴趣区域。
41	tmRNA	×	√	转移信使 RNA；tmRNA 首先充当 tRNA，然后充当编码肽标签的 mRNA；核糖体翻译 tmRNA 的该 mRNA 区域并将编码的肽标签连接到未完成的蛋白质的 C 末端；该附着的标签靶向蛋白质用于破坏或蛋白水解。
42	transit_peptide	√	√	
43	tRNA	√	√	
44	unsure	√	√	
45	V_region	√	√	
46	V_segment	√	√	
47	variation	√	√	
48	3'UTR	√	√	
49	5'UTR	√	√	
50	allele	√	×	ST.26 中 allele 为 misc_feature 特征关键词的限定词
51	attenuator	√	×	ST.26 中 attenuator 为 regulatory 特征关键词的限定词 regulatory_class 的取值
52	CAAT_signal	√	×	ST.26 中 CAAT_signal 为 regulatory 特征关键词的限定词 regulatory_class

				的取值
53	conflict	√	×	ST.26 中 allele 为 misc_feature 特征关键词的限定词 note 的取值
54	enhancer	√	×	ST.26 中 enhancer 为 regulatory 特征关键词的限定词 regulatory_class 的取值
55	GC_signal	√	×	ST.26 中 GC_signal 为 regulatory 特征关键词的限定词 regulatory_class 的取值
56	LTR	√	×	ST.26 中 LTR 为 mobile_element 特征关键词的限定词 rpt_type 的取值
57	misc_signal	√	×	ST.26 中 misc_signal 为 regulatory 特征关键词的限定词 regulatory_class 的取值(other)
58	Mutation	√	×	ST.26 中 mutation 为 variation 特征关键词的限定词 note 的取值
59	old_sequence	√	×	ST.26 中 old_sequence 为 misc_feature 特征关键词的限定词 note 的取值
60	polyA_signal	√	×	ST.26 中 polyA_signal 为 regulatory 特征关键词的限定词 regulatory_class 的取值(polyA_signal_sequence)
61	promoter	√	×	ST.26 中 promotor 为 regulatory 特征关键词的限定词 regulatory_class 的取值
62	RBS	√	×	ST.26 中 RBS 为 regulatory 特征关键词的限定词 regulatory_class 的取值(ribpspme_binding_site)
63	repeat_unit (when repeat _region not used)	√	×	ST.26 中 repat_unit 为 misc_feature 特征关键词的限定词 note 的取值
64	repeat_unit (when repeat _region used)	√	×	ST.26 中 repat_unit 为 repeat_region 特征关键词的限定词 rpt_unit_range 的取值
65	satellite	√	×	ST.26 中 satellite 为 repeat_region 特征关键词的限定词
66	scRNA	√	×	ST.26 中 scRNA 为 ncRNA 特征关键词的限定词 ncRNA_class 的取值
67	snRNA	√	×	ST.26 中 snRNA 为 ncRNA 特征关键词的限定词 ncRNA_class 的取值
68	TATA_signal	√	×	ST.26 中 TATA_box 为 regulatory 特征关键词的限定词 regulatory_class 的取值
69	terminator	√	×	ST.26 中 terminator 为 regulatory 特征关键词的限定词 regulatory_class 的取值
70	3'clip	√	×	ST.26 中 3'clip 为 misc_feature 特征关键词的限定词 note 的取值
71	5'clip	√	×	ST.26 中 5'clip 为 misc_feature 特征关键词的限定词 note 的取值

72	-10_signal	√	×	ST.26 中 minus_10_signal 为 regulatory 特征关键词的限定词 regulatory_class 的取值
73	-35_signal	√	×	ST.26 中 minus_35_signal 为 regulatory 特征关键词的限定词 regulatory_class 的取值

(2) 氨基酸序列的特征关键词表的比较

ST.25 序列表中为蛋白质序列表的特征关键词表，ST.26 中为氨基酸序列表的特征关键词表。

表 10 给出了 ST.25 和 ST.26 氨基酸序列的特征关键词对照表，从表 10 可以看出：

ST.25 中有，ST.26 中没有的氨基酸序列特征关键词共 20 个，主要区别是 ST.25 的特征关键词在 ST.26 中为某个特征关键词的限定词的取值。ST.25 中没有，ST.26 中有的氨基酸序列特征关键词共 10 个，分别是 COILED、COMPBIAS、CROSSLNK、INTRAMEM、MOTIF、NON_STD、REGION、SOURCE、TOPO_DOM 和 VAR_SEQ。

ST.25 序列表的特征关键词表包括“关键词”和对关键词的“说明”，ST.26 序列表的特征关键词表中包括关键词、关键词定义、可选限定词、(强制性限定词词、生物体范围、分子范围、补充说明等）内容，其中关键词、关键词定义、可选限定词为所有关键词都出现的项，其他项不是所有关键词都有相应的信息。

表 10 ST.25 和 ST.26 氨基酸序列的特征关键词对照表

序号	蛋白质/氨基酸序列的特征关键词表	ST.25	ST.26	备注
1	ACT_SITE	√	√	
2	BINDING	√	√	
3	CA_BIND	√	√	
4	CARBOHYD	√	√	
5	CHAIN	√	√	
6	COILED	×	√	卷曲螺旋区域的范围
7	COMPBIAS	×	√	成分偏置区域的范围

8	CONFLICT	√	√	
9	CROSSLNK	×	√	翻译后形成的氨基酸键
10	DISULFID	√	√	
11	DNA_BIND	√	√	
12	DOMAIN	√	√	
13	HELIX	√	√	
14	INIT_MET	√	√	
15	INTRAMEM	×	√	位于膜中但不与其交叉的区域的范围
16	LIPID	√	√	
17	METAL	√	√	
18	MOD_RES	√	√	
19	MOTIF	×	√	生物学感兴趣的短（最多 20 个氨基酸）序列基序
20	MUTAGEN	√	√	
21	NON_STD	×	√	非标准氨基酸
22	NON_TER	√	√	
23	NP_BIND	√	√	
24	PEPTIDE	√	√	
25	PROPEP	√	√	
26	REGION	×	√	序列中感兴趣区域的范围
27	REPEAT	√	√	
28	SIGNAL	√	√	
29	SITE	√	√	
30	SOURCE	×	√	标识序列的来源；该关键词是强制性的；每个序列将具有跨越整个序列的单个源特征
31	STRAND	√	√	
32	TOPO_DOM	×	√	拓扑域
33	TRANSMEM	√	√	

34	TRANSIT	√	√	
35	TURN	√	√	
36	UNSURE	√	√	
37	VARIANT	√	√	
38	VAR_SEQ	×	√	通过可变剪接、可变启动子使用、可变起始和核糖体移码产生的序列变体的描述
39	ZN_FING	√	√	
40	VARSPLIC	√	×	ST.26 中 VARSPLIC 为 VAR_SEQ 特征关键词的限定词 NOTE 的取值
41	ACETYLATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
42	AMIDATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
43	BLOCKED	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
44	FORMYLATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
45	GAMMA-CARBOXYGLUTAMIC ACIDHYDROXYLATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
46	METHYLATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
47	PHOSPHORYLATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
48	PYRROLIDONECARBOXYLIC ACID	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
49	SULFATATION	√	×	ST.26 中 MOD_RES 特征关键词的限定词 NOTE 的取值
50	MYRISTATE	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
51	PALMITATE	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
52	FARNESYL	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
53	GERANYL-GERANYL	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
54	GPI-ANCHOR	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
56	N-ACYLDIGLYCERIDE	√	×	ST.26 中 LIPID 特征关键词的限定词 NOTE 的取值
57	THIOLEST	√	×	ST.26 中 CROSSLINK 特征关键词的限定词 NOTE 的取值

58	THIOETH	√	×	ST.26 中 CROSSLINK 特征关键词的限定词 NOTE 的取值
59	SIMILAR	√	×	ST.26 中 REGION 特征关键词的限定词 NOTE 的取值
60	NON_CONS	√	×	ST.26 中 SITE 特征关键词的限定词 NOTE 的取值

4. 特征限定词的比较

ST.25 没有特征限定词。ST.26 中的特征限定词用于提供除了由特征关键词和特征位置传达的特征之外的信息。限定词有三种类型的值格式（自由文本[①]、受控词表或枚举值、序列）来传达的不同类型的信息。作为限定词值提供的任何序列必须单独列入序列表并分配其自己的序列标识号。一个强制性特征关键词需要两个强制性限定词，一些非强制性特征关键词也需要强制性限定词。

(1) 核苷酸序列的限定词表的比较

ST.26 的限定词表包括限定词、限定词定义、取值形式、（举例、补充说明）等内容，其中限定词、限定词定义、取值形式为所有限定词均出现的项，其他项不是所有限定词都有相应的信息。

ST.26 的限定词表中共有 80 个限定词（详见表 11），除 allele、gene、satellite 为 ST.25 中核苷酸序列的特征关键词外，其余均为新增的限定词。

表 11 ST.26 标准中核苷酸序列的限定词表

序号	核苷酸序列的限定词表	备注
1	allele	在 ST.25 中，allele 为核苷酸序列的特征关键词
2	anticodon	
3	bound_moiety	
4	cell_line	
5	cell_type	
6	chromosome	
7	clone	
8	clone_lib	
9	codon_start	
10	collected_by	
11	collection_date	
12	compare	

① ST.25 中“自由文本”要求每个数据元素的字符数不超过 4 行，每行不超过 65 个字符；ST.26 中的“自由文本”不超过 1000 字符。

13	cultivar	
14	dev_stage	
15	direction	
16	EC_number	
17	ecotype	
18	environmental_sample	
19	exception	
20	frequency	
21	function	
22	gene	在 ST.25 中，gene 为核苷酸序列的特征关键词
23	gene_synonym	
24	germline	
25	haplogroup	
26	haplotype	
27	host	
28	identified_by	
29	isolate	
30	isolation_source	
31	lab_host	
32	lat_lon	
33	macronuclear	
34	map	
35	mating_type	
36	mobile_element_type	
37	mod_base	
38	mol_type	
39	ncRNA_class	
40	note	
41	number	
42	operon	
43	organelle	
44	organism	
45	PCR_primers	
46	phenotype	
47	plasmid	
48	pop_variant	
49	product	
50	protein_id	
51	proviral	
52	pseudo	
53	pseudogene	
54	rearranged	
55	recombination_class	

56	regulatory_class	
57	replace	
58	ribosomal_slippage	
59	rpt_family	
60	rpt_type	
61	rpt_unit_range	
62	rpt_unit_seq	
63	satellite	在 ST.25 中，satellite 为核苷酸序列的特征关键词
64	segment	
65	serotype	
66	serovar	
67	sex	
68	standard_name	
69	strain	
70	sub_clone	
71	sub_species	
72	sub_strain	
73	tag_peptide	
74	tissue_lib	
75	tissue_type	
76	transl_except	
77	transl_table	可以与“CDS”特征关键词一起使用以指示该区域将使用替代遗传码表来翻译。
78	trans_splicing	
79	translation	
80	variety	

(2) 氨基酸序列的限定词表的比较

ST.26 中的氨基酸序列的限定词有 MOL_TYPE、NOTE 和 ORGANISM 三个（见表 12），每个限定词包括限定词定义、取值形式、举例和补充说明。

表 12 ST.26 标准中核苷酸序列的限定词表

序号	氨基酸序列的限定词表	备注
1	MOL_TYPE	SOURCE 特征关键词强制使用 MOL_TYPE 限定词。
2	NOTE	对特征关键词 BINDING、CARBOHYD、CROSSLINK、DISULFID、DOMAIN、LIPID、METAL、MOL_RES、NP_BIND 和 ZN_FING，NOTE 是强制性限定词。
3	ORGANISM	对 SOURCE 特征关键词 ORGANISM 是强制性限定词。

5. 基因编码表的比较

基因编码表主要用于翻译编码序列。ST.25 中不提供标准化方式表明将使用除标准基因编码表之外的基因编码表来翻译核苷酸序列的 CDS 区。ST.26 中的限定词“transl_table”，可以与“CDS”特征关键词一起使用以表明将使用替代基因编码表来翻译该区域。如果不使用限定词“transl_table”，那么假定使用标准基因编码表。

6. 特征位置的比较

ST.25 中没有提供用来表明特征位置的标准化方式，特别是包含在延伸超出一个特定残基或多个残基跨度的位点或区域中的特征，例如延伸超出所公开序列的一端或两端的核苷酸序列的 CDS 区域。ST.26 中的特征位置描述符（“<”或“>”）提供了一种标准化方式用来表明这样的位点或区域的位置。

三、 标准变化对智能化升级系统设计的影响

核苷酸和氨基酸序列表的表述标准从 ST.25 转变为 ST.26，在适用范围、文件格式和标准内容等方面都有所变化。在进行智能化系统设计时，需要考虑这些变化对系统设计的影响，主要体现在应用系统设计、数据库设计和历史数据处理三个方面。

(一) 应用系统设计

1. 电子申请系统

ST.26 标准适用于国际、国内或区域程序的专利申请，因此在国内专利和国际专利的新一代智能电子申请系统的设计中，都需要提供符合 ST.26 标准的序列表编辑和校验工具。

- 新一代智能电子申请系统的智能客户端，需要提供序列表生成、导入、编辑和验证功能的模块。
  - 提供符合 ST.26 标准的序列表生成和编辑工具，申请人可以新建序列表或编辑已有的序列表；
  - 提供序列表导入和批量导入功能，导入文件后定位到编辑界面；
  - 提供序列表校验工具，对编辑完成后的序列表进行格式校验，保

证文件内容可以正常提取，经过校验的序列表保存为符合 ST.26 标准的 XML 文件。

此外，与 ST.25 标准不同，ST.26 标准以特征关键词和特征限定词共同传递序列信息，且对某些特征关键词，必须使用强制性限定词进行表述（如在核苷酸列表的特征关键词 source 需使用其强制性限定词 organism 和 mol_type 表述）。因此，在进行编辑和校验工具设计时，应考虑此类强制规则，在编辑序列表时进行智能填充，在校验序列表时给出智能提示。

- 新一代智能电子申请系统的智能交互式平台，需提供符合 ST.26 标准的序列表导入、在线编辑和在线校验工具，具体功能应与电子申请智能客户端相同。

2. **智能审查系统**

ST.26 标准使用 XML 文件表示序列表，因此在新一代智能审查系统的设计中，应当考虑对 XML 文件进行处理，并从 ST.26 标准中提取统一规则，构建统一的 XML 解析模块，获取序列表中的著录项目数据和序列内容数据。根据业务需求，将这些规则和数据应用在不同审查模块的具体功能中。

- 新一代智能审查系统的受理模块需要支持符合 ST.26 标准的序列表的校验功能。特别是对随纸件申请提交的计算机可读形式的电子序列表，需经过基本的格式校验，符合 ST.26 标准的要求才能进入审查系统。
- 新一代智能审查系统的初审模块需要支持符合 ST.26 标准的序列表的形式审查功能。对电子序列表内容，首先审查其著录项目信息是否与专利申请相一致，如发明名称、申请人等是否与请求书中相同；其次，形式审查还应包括对 XML 文件中数据项是否符合 ST.26 标准要求的审查，如受理局是否使用了 ST.3 标准规定的双字母表示，申请日等日期格式是否正确，序列特征的位置描述是否与标准要求一致等。
- 新一代智能审查系统的实审模块需要支持对 XML 格式的电子序列表的展示功能。将序列表内容以适合审查员阅读的方式展示在系统

界面上，同时转换序列格式以便推送给检索系统进行智能检索。

3. 智能检索系统

新一代智能检索系统接收审查系统推送的生物序列，整合多个生物序列数据库入口，提供统一的序列检索平台进行检索，并提供序列比对功能。

- 在 ST.25 标准中，氨基酸用首字母大写的三字母代码表示，而在 ST.26 标准中，氨基酸用单个大写字母表示。目前检索的各生物序列数据库均需使用单字母表示的序列，与 ST.26 标准的要求一致，对符合 ST.26 标准的生物序列表，不需要特别考虑氨基酸序列的三字母与单字母转换。但对于 ST.26 标准实施之前提交的生物序列表，因申请人可能提交三字母形式的氨基酸序列，因此在进行智能检索系统设计时，仍需要提供三字母形式的氨基酸序列到单字母氨基酸序列的转换功能。
- ST.26 标准提供了遗传密码表，用于对三个核苷酸碱基和一个氨基酸残基进行对照翻译。申请人可能仅提交核苷酸或氨基酸序列，而在检索时应当将其拓展到对应的氨基酸或核苷酸序列，因此在进行智能检索系统设计时，需要提供按照 ST.26 标准的遗传密码表进行序列翻译的功能。

(二) 数据库设计

ST.26 标准采用 XML 格式，相对于 ST.25 标准使用的 TXT 文件，XML 文件在数据解析和管理方面更加安全和快捷。在设计新一代智能化系统的数据库时，可以考虑设置相应的生物序列数据表，以使生物序列信息更好的被系统记录和使用。

生物序列数据表在数据字段的设置上，需包含 ST.26 标准给出的 DTD 规范中的重要数据项。在数据字段属性的设置上，也需满足 ST.26 标准的要求，如字段是否可为空，字段长度要求等，都要与标准要求相一致。

(三) 历史数据处理

ST.26 标准将于 2022 年开始全面实施，在此之前提交的生物序列表，仍

将是依据 ST.25 标准或中国生物序列标准生成的。在新一代智能化系统中，如何处理历史数据是一个需要重点考虑的问题。

ST.26 标准的附件 7 对如何将序列表从 ST.25 转换成 ST.26 给出了详细的建议。理论上，根据 ST.25 和 ST.26 的数据对照关系和附件 7 的指导，可以将所有的依据 ST.25 标准提交的历史文件转换为符合 ST.26 标准的文件。因此，标准变化在历史数据处理方面对系统设计的影响主要在于，需决定在何时、以何种形式对历史数据进行转换，使新一代智能化系统对生物序列的展示、审查和检索等功能不受历史数据的影响。例如，可以在系统上线前对历史数据进行批量转换，将所有历史文件转换为符合 ST.26 标准的 XML 文件，后续系统功能只需处理 ST.26 格式的序列表；或保留历史文件，在系统中设计数据转换模块，在系统需要对历史文件进行展示、审查和检索时，实时进行文件转换等。历史数据处理方案需根据历史文件数量、批量转换成本、模块开发成本等因素进行具体分析后作出选择。

第四章 中国生物序列标准概述

一、 中国生物序列标准的现状

(一) 标准简介

2001 年 11 月，国家知识产权局发布并实施了行业标准——《核苷酸和/或氨基酸序列表和序列表电子文件标准》（ZC0003-2001），该标准主要是参考 WIPO 的 ST.25 标准（1998 版）并结合中国实际工作情况制定的[①]，在内容上与 WIPO 的 ST.25 标准保持一致。2017 年，国家知识产权局对标准中的部分表示方式进行了的修订[②]，但实质内容未做改动。

(二) 标准的目的

标准制定的目的是为了使以纸件形式提交的核苷酸和/或氨基酸序列表及计算机可读形式的含有该序列表的电子文件规范化，以利于申请人提交；也为了使序列表的电子文件可以快捷地输入国家知识产权局专利局的计算机数据库，并与其它的序列检索数据库交换数据，以利于公众检索；同时也利于专利局审查员加快审查，更好地为申请人服务。

(三) 标准的适用范围

《核苷酸和/或氨基酸序列表和序列表电子文件标准》（ZC0003-2001）作为中华人民共和国知识产权行业标准，适用于所有向国家知识产权局专利局提交的包含核苷酸和/或氨基酸序列的发明专利申请，具体地说，适用于以纸件形式提交的核苷酸和/或氨基酸序列表，以及含有核苷酸和/或氨基酸序列表的计算机可读形式的序列表电子文件。

二、 中国生物序列标准的使用情况

(一) 法律层面的体现

①曲超，张松，曲晓光. WIPO 标准 ST.26 与 ST.25 对比研究分析[J]. 中国标准化, 2012(9):50-52.
②关于知识产权行业标准《电子文件标准》中部分特征关键词表的修订（第 248 号）.(2017-07-26) [2019-07-17] http://www.sipo.gov.cn/docs/pub/old/zwgg/gg/201707/t20170726_1312897.html

《核苷酸和/或氨基酸序列表和序列表电子文件标准》（ZC 0003-2001）的强制性主要体现在《中华人民共和国专利法实施细则》及《专利审查指南》中相应条款的规定，主要包括：

1. 《专利法实施细则》中关于生物序列的规定

《专利法实施细则》第 17 条第 4 款的规定，发明专利申请包含一个或者多个核苷酸或者氨基酸序列的，说明书应当包括符合国务院专利行政部门规定的序列表。申请人应当将该序列表作为说明书的一个单独部分提交，并按照国务院专利行政部门的规定提交该序列表的计算机可读形式的副本。

2. 《专利审查指南》中的相关规定

(1) 受理程序

《专利审查指南》第五部分第三章第 2.3.1 节受理程序规定，（ 2 ）对于涉及核苷酸或者氨基酸序列的发明专利申请，还应当核实是否提交了包含相应序列表的计算机可读形式的副本，例如光盘或者软盘等。

(2) 初步审查

《专利审查指南》第一部分第一章第 4.2 节规定，涉及核苷酸或者氨基酸序列的申请，应当将该序列表作为说明书的一个单独部分，并单独编写页码。申请人应当在申请的同时提交与该序列表相一致的计算机可读形式的副本，如提交记载有该序列表的符合规定的光盘或者软盘。提交的光盘或者软盘中记载的序列表与说明书中的序列表不一致的，以说明书中的序列表为准。未提交计算机可读形式的副本，或者所提交的副本与说明书中的序列表明显不一致的，审查员应当发出补正通知书，通知申请人在指定期限内补交正确的副本。期满未补交的，审查员应当发出视为撤回通知书。

(3) 实质审查

《专利审查指南》第二部分第八章第 4.7.2 节，审查说明书和摘要中规定，专利申请包含一个或多个核苷酸或氨基酸序列的，应当审查说明书是否包括符合规定的序列表。

《专利审查指南》第二部分第十章第 9.2.3 节，核苷酸或氨基酸序列表（1）当发明涉及由 10 个或更多核苷酸组成的核苷酸序列，或由 4 个或更多Ｌ－氨基酸组成的蛋白质或肽的氨基酸序列时，应当递交根据国家知识产权

局发布的《核苷酸和/或氨基酸序列表和序列表电子文件标准》撰写的序列表。序列表应作为单独部分来描述并置于说明书的最后。此外申请人还应当提交记载有核苷酸或氨基酸序列表的计算机可读形式的副本。如果申请人提交的计算机可读形式的核苷酸或氨基酸序列表与说明书和权利要求书中书面记载的序列表不一致，则以书面提交的序列表为准[①]。

(二) 用户层面的体现

1. 申请阶段的使用

对于电子申请，一般情况下，国内专利申请、PCT 申请的国家阶段、巴黎公约发明的电子申请是通过电子申请客户端（CPC）递交，PCT 国际申请通过 CEPCT 电子申请客户端提交。CPC 客户端集成有生物序列表在线编辑和校验功能，校验标准依照《核苷酸和/或氨基酸序列表和序列表电子文件标准》（ZC0003-2001）；目前的 CEPCT 电子申请客户端没有集成序列表编辑和校验功能，CEPCT 电子申请可以接受 APP 格式和 TXT 格式的生物序列表。

对于纸件申请，在递交纸件的序列表的同时提交、主动补正或是官方下发补正书后补交 TXT 格式的序列表，对序列表是否符合标准由申请人自己把控，但是常用的专利序列表生成软件，如国家知识产权局提供的 SIPOSequenceListing，欧洲专利局提供的 patentin，生成的序列表符合 ST.25 标准。

2. 流程审查阶段的使用

对于国内申请，在受理环节的服务端做相关校验；在初审环节，发明智能审查系统中有针对序列表的审查规则，该规则调用受理环节配置的校验信息；在实审环节，E 系统实审子系统提供查看序列表信息和内容的功能以及下载序列表 TXT 文件的功能。

对于国外申请，在流程审查阶段有校验环节，有专人负责校验工作，国际阶段审查过程中，有独立的校验软件供审查员使用。

① 《专利审查指南 2010》[EB/OL].[2019-07-17]. http://www.sipo.gov.cn/zhfwpt/zlsqzn_pt/zlfssxzjsczn/index.htm.

三、 中国生物序列加工情况分析

(一) 项目概况及意义

中国生物序列的加工主要是知识产权出版社承接的“中国专利生物序列数据标引”服务合同，该加工项目的目的是为了使提交的中国专利中的核苷酸和/或氨基酸序列便于检索，与其他的序列检索数据库交换数据。该项目是参照世界知识产权组织（WIPO）ST.25 标准和中华人民共和中国知识产权行业标准《核苷酸和/或氨基酸序列表电子文件标准》（ZC0001-2003），将中国专利文献中的生物序列信息（核苷酸和/或氨基酸序列信息）加工成标准的电子序列表文件。

(二) 具体实施方式

技术服务的主要内容包括在规定的时间内完成某一时间段内的中国专利发明公开数据中包含生物序列专利的筛选及生物序列的标引和翻译工作，具体加工方式包括：数据筛选、著录项目标引、专利文献中核苷酸和/或氨基酸序列及其相关信息标引、文件生成、翻译、数据整合等。标引过程采用出版社自主开发的“专利生物序列表生成软件”处理专利说明书的著录项目和生物序列信息。“专利生物序列表生成软件”包括序列编辑器和序列生成器：序列编辑器中可以输入、修改核酸或蛋白质序列，也可以从其他编辑器或者 word(保存为 ASCII 的文本文件)导入序列表文档;序列生成器是在输入专利申请所需的所有数据后，生成计算机可读的并符合 ST.25 标准的序列表文件[①]。

(三) 为标准过渡所做的准备

2015 年 10 月，由国知局条法司、专利局审查业务管理部、初审及流程管理部、医药生物发明审查部、专利文献部、自动化部，知识产权出版社等相关部门人员组成项目组，对照当时的 ST.26 标准草案[②]制定了符合 ST.26

①中国专利生物序列数据标引技术服务合同(2017-2018)[EB/OL]. [2019-06-20] http://www.sipo.gov.cn/ztzl/zfcgzl/htgs/1124874.htm.

② ST.26 标准于 2016 年 3 月正式通过，后又经历了几次修订。

标准的数据加工规则、加工方式。知识产权出版社根据 WIPO 的 ST.26 标准和数据加工规则开发了符合 WIPO 的 ST.26 标准的序列标引流程管理系统和加工软件，以及将符合 WIPO 的 ST.25 标准的序列数据转换为符合 ST.26 标准的序列数据转换软件。

第五章 ST.26 标准配套软件工具的评估

一、 ST.26 标准配套软件工具概述

2017 年 5 月，在 CWS 第五届会议上，标准委员会商定所有知识产权局同时从 ST.25 向 ST.26 过渡，即 2022 年 1 月 1 日之后收到的所有申请都必须符合 ST.26 标准；国际局通知标准委员会，将开发一种通用软件工具，使申请人能够编制序列表并验证此种序列表是否符合 ST.26 标准。

2019 年 3 月，国际局正式批准了所开发的通用工具及其代表组件的名称为“WIPO Sequence 工具”。WIPO Sequence 工具界面将以 PCT 的十种正式公布语言提供（英文、阿拉伯文、中文、法文、德文、日文、韩文、葡萄牙文、俄文和西班牙文），在工具首次发布时即以上述各种语言提供，但为申请人提供有关如何使用该工具基本支出的“用户指南”此次仅以英文提供。

(一) 工具的组成

“WIPO Sequence 工具”由三个组件组成，分别是：

- WIPO Sequence：由申请人本地安装的桌面应用程序，可供申请人编著和验证其想要寻求专利保护的序列表；
- WIPO Sequence Validator：集成到知识产权局环境中的微服务，确保各知识产权局仅接受符合要求的序列表；
- WIPO Sequence Server：国际局用于提供该工具新版本的更新和发布服务器。

(二) 工具的特点

WIPO Sequence 是一个独立的应用程序，安装在本地电脑上，由申请人使用。

WIPO Sequence Validator 作为微服务部署在服务器环境中，与知识产权局使用的其他业务解决方案的应用程序通信，以便对申请人提交序列表数据提供验证服务。

WIPO Sequence Server 部署在产权组织网络中，并提供新版本的 WIPO Sequence 和/或 WIPO Sequence Validator。

(三) 工具的功能

WIPO Sequence 的主要功能包括(1)从用户获取数据，并创建 ST.26XML 格式的序列表文件；（2）核实一个序列表是否符合 ST.26 标准的要求；（3）从 ST.25、ST.26 等各种格式和其他行业格式的外部文件中导入数据，必要时从用户处获取更多输入信息，以便生成 ST.26 符合标准的 XML 序列表。

WIPO Sequence Validator 的主要功能是验证所提交的序列表是否符合 ST.26 标准的要求。

WIPO Sequence Server 的主要功能是提供 WIPO Sequence 和/或 WIPO Sequence Validator 的更新版本。

二、 ST.26 标准配套软件工具在智能化升级系统中应用的可行性评估

(一) 系统和硬件评估

1. WIPO Sequence 的系统及硬件要求

WIPO Sequence 在计算机上安装应用程序需要权限。

(1) 对系统的要求

WIPO Sequence 工具将在操作系统中获得认证：

- Windows 10 1803 版
- Ubuntu 18.04 版
- MacOS 10.13 版（64 位版本）
- CentoS 7 1804 版

除了认证应用程序的版本之外，申请人工具还应当可以在以下操作系统上使用，因为该工具的核心组件得到了支持：

- Windows 7 及更高版本（32 位和 64 位）
- Ubuntu 12.04 版及更新版本

- MacOS 10.9 版（64 位版本）
- Debian 8

(2) 对硬件的要求

申请人工具达到以下最低硬件要求将获得认证：

- 中央处理器：1.6GHz
- 内存：4GB
- 可用硬盘：1GB（存储序列表信息可能需要更大的硬盘）
- 屏幕分辨率：1366x768

2. WIPO Sequence Validator 的系统及硬件要求

(1) 对系统的要求

知识产权局工具将基于 SpringBoot2.0.3，并且需要支持以下基本软件组件的操作系统：

- Java8
- 小服务程序 3.1 容器。（Tomcat 8.5 将用作默认的小服务程序容器）

(2) 对硬件的要求

知识产权局工具达到以下最低硬件要求将获得认证：

- 中央处理器：1.6GHz
- 内存：4GB
- 可用硬盘：1GB（存储序列表信息可能需要更大的硬盘）

(二) WIPO Sequence 的测试与评估

根据 WIPO Sequence 的用户手册和测试模板对 WIPO Sequence 进行了测试与评估。

1. 测试环境

操作系统	Windows 7 专业版
接口语言	中文

2. 测试内容

依据 WIPO Sequence 测试模板所提供的功能检查表（见表 13），对各项功能逐一进行了测试。

表 13 WIPO Sequence 功能检查表

序号	是否通过（Y/N）	功能	评论
1	Y	将自定义有机体名称添加到系统的有机体名称列表中。	同一生物体（文件）重复创建/导入时，系统不会重复显示，但不会提示。
2	Y	将发明名称及其对应的语言代码添加到项目中。	
3	Y	将申请信息（当前申请或以前的申请）添加到项目中。	
4	Y	将特征信息添加到序列中。	
5	Y	将源特征及其强制限定词添加到序列中。	
6	Y	将新的有机体名称添加到存储在此系统中的有机体名称列表中。	
7	Y	向特征添加限定词信息。	
8	Y	将序列表的通用信息数据添加到项目中；	
9	Y	更改序列将在生成的序列表中列出的顺序。	
10	Y	创建一个工作区，其中存储与一个序列表相关的数据。	用户手册和 WIPO SequenceTool 测试工具中均未出现“workspace”；如果 workspace 是指 project，则此项通过测试。
11	Y	创建序列数据结构的实例，并将其属性设置为从作为输入接收的 ST.26 序列数据 XML 节点获得的值。	
12	Y	创建序列并将其插入列表中的其他位置。	
13	Y	生成序列表。	
14	Y	显示生成的序列表。	
15	Y	编辑项目的属性。	
16	Y	编辑序列的属性。	
17	Y	编辑特征数据结构实例的属性。	如果 attributes 是指特征的名称和取值的情况下，则此项测试通过。
18	Y	编辑限定词数据结构实例的属性。	如果 attributes 是指限定词的名称和取

			值的情况下，则此项测试通过。
19	Y	删除序列。	
20	N	启用或禁用选定的验证规则。	未发现验证规则设置选项。
21	Y	创建新的个人或组织名称。	
22	Y	导出存储在项目中的所有数据，以便以后可以将其导入到系统的相同或不同实例中。	
23	Y	将自定义有机体名称列表导出到一个文件中，该文件稍后可以导入到此系统的另一个实例中。	
24	Y	从文件导入自定义有机体名称列表。	
25	N	导入存储在项目文件中的所有数据（例如，从其他律师事务所）。	不支持包含一个或多个 ST.25 格式生物序列 txt 文件的压缩包 (.zip) 的项目导入。
26	Y	将 ST.25 序列表文件中的数据导入到新创建的项目中。	
27	Y	将 ST.26 序列表文件中的数据导入到新创建的项目中。	
28	Y	一次导入文件中的多个序列。	
29	Y	将来自另一个项目（源项目）的数据导入当前项目（目标项目）。	导入后，通用信息部分被重写。
30	Y	打印项目数据或生成 ST.26 序列表。	
31	Y	向所选特征提供位置信息。	
32	Y	从项目中删除与序列相关联的所有数据，并相应地对其余序列重新编号。	
33	Y	在系统中存储关于申请人或发明人的信息（例如，姓名、其相应的语言代码及其翻译或音译为拉丁文字符（如果适用）、地址等），以便以后在各种项目中使用。	后添加的申请人/发明人将会替换先添加的申请人/发明人。
34	Y	验证 ST.26 序列表文件，并将问题作为包含警告和错误消息的验证报告列出；	
35	Y	验证存储在项目中的数据，并将问题作为包含警告和错误消息的验证报告列出。	
36	Y	为翻译目的输出自由文本限定词。	
37	Y	向系统提供包含无效符号的残基字符串，并验证重新格式化的进程残基。	提示输入的残基符号无效。
38	Y	为选定的 CDS 功能及其关联的翻译序列创建翻译限定词。	

39	Y	根据指定的遗传密码表号翻译核酸序列。	
40	N	解析一个多序列格式的序列并检查它返回 4 个部分（序列名、分子类型、有机体和残基）。	
41	Y	记录基于导入更改的数据，以便导入后可以显示原始数据和更改后的数据。	
42	Y	显示导入序列时更改的数据。	
43	Y	将序列名添加到所选序列。	
44	Y	将序列的 INSDQualifierMolType 属性设置为预定义值之一。	
45	Y	设置并存储系统偏好（如每行显示的最大残基符号数等）。	

3. 缺陷报告

WIPO Sequence 的测试模板对缺陷和优先级进行了定义[①]，具体如下：

（1） 严重性

严重性是缺陷可能影响软件的程度，定义了给定缺陷对系统的影响。

按照缺陷对系统影响的强度，严重性划分为 Bocker、Critical、Major 和 Minor 四个等级。

（2） 优先级

优先级定义了开发团队解决缺陷的顺序，优先级状态是根据产品所有者的要求设置的。

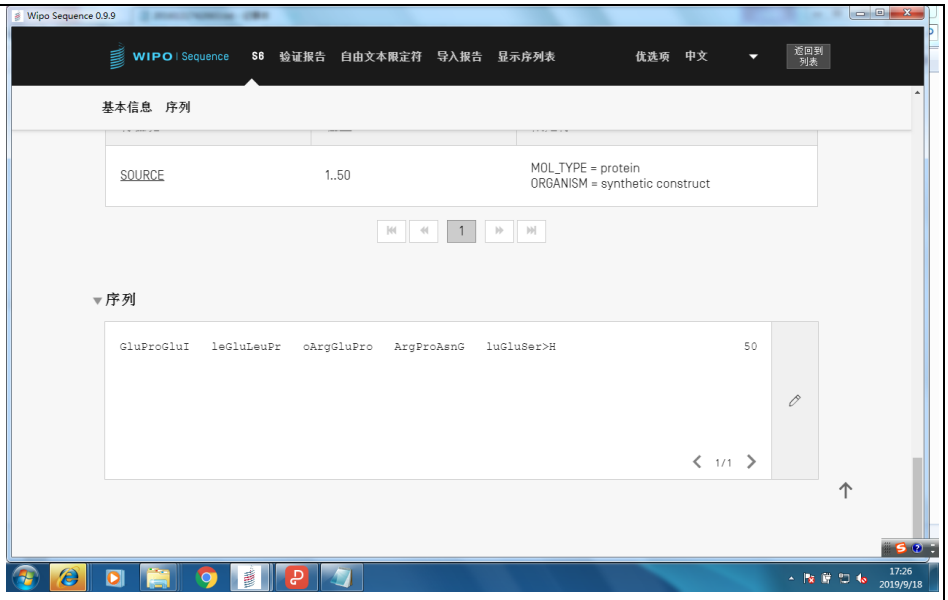
按照缺陷需解决的顺序，优先级分为 High、Medium 和 Low 三个等级。表 14 列出了 WIPO Sequence 测试的缺陷报告。

表 14 WIPO Sequence 缺陷报告

序号	字段	说明
1	说明	申请人/发明人只能添加一次，后添加的申请人/发明人将会替换先添加的申请人/发明人
	组件	WIPO Sequence Tool
	环境	Windows 7；Chinese interface
	序列数	
	严重性	Minor
	优先级	Low
	复制步骤	
	实际结果	后添加的申请人/发明人将会替换先添加的申请人/发明人；本轮测试中，

① 详见 WIPO Sequence Formal UAT Report。

		对多个申请人/发明人用“,”隔开；但系统没有说明。
	预期结果	建议允许添加多个申请人/发明人，可以创建申请人/发明人列表，或者对于多个申请人/发明人的添加规则，系统给予说明。
2	说明	项目名称可以重复
	组件	WIPO Sequence Tool
	环境	Windows 7；Chinese interface
	序列数	
	严重性	Minor
	优先级	Low
	复制步骤	为多个项目使用同一项目名称： 
	实际结果	允许多个项目使用同一名称
	预期结果	当项目名称重复时，建议提示名称已存在，改用新的项目名称。
3	说明	ST.25 的蛋白质序列导入时，会出现“三字母”不能转换成“单字母”的情形。
	组件	WIPO Sequence Tool
	环境	Windows 7；Chinese interface
	序列数	23
	严重性	Major
	优先级	Medium
	复制步骤	（1）导入多序列格式的符合 ST.25 标准的 txt 文件；（2）序列 13（PRT）的“三字母”代码未能正确转换。

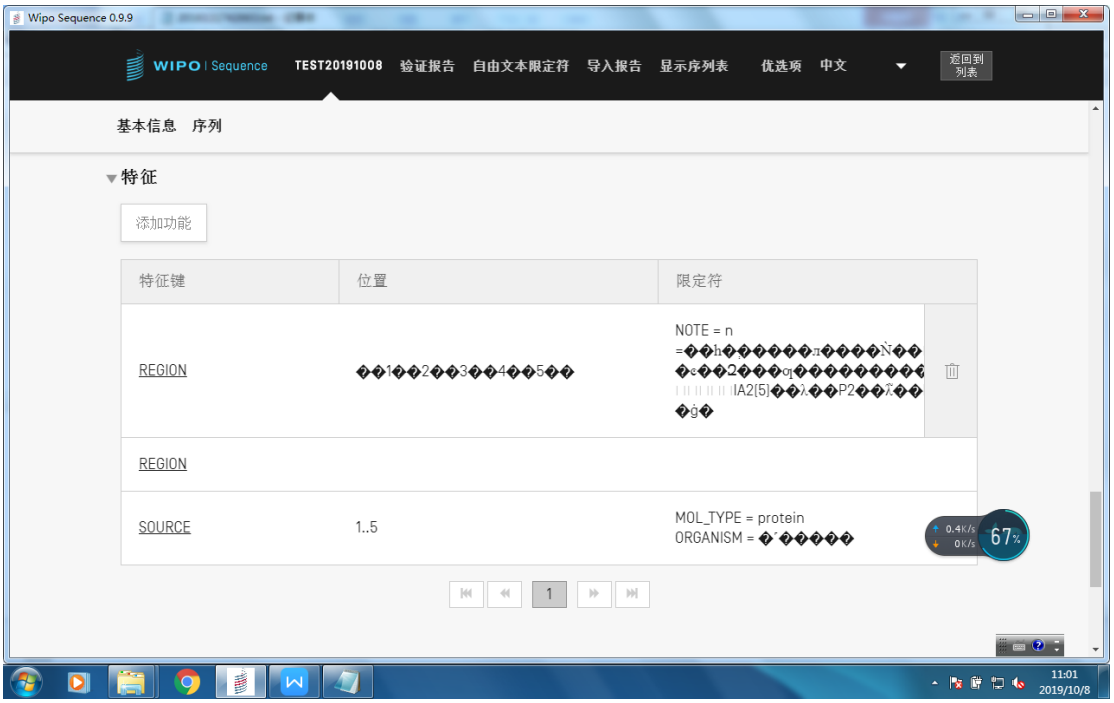
	实际结果	
	预期结果	系统自动将“三字母”转换成对应的“单字母”代码。
4	说明	不支持 ST.25 格式生物序列的.zip 项目导入
	组件	WIPO Sequence Tool
	环境	Windows 7；Chinese interface
	序列数	
	严重性	Major
	优先级	Medium
	复制步骤	（1）导入项目（2）选中 ST.25 格式的生物序列的.zip 文件 
	实际结果	导入项目，不能导入 ST.25 格式的生物序列文件
	预期结果	建议项目导入增加对 ST.25 格式生物序列文件的支持，并且系统能够说明导入项目的.zip 文件需要满足的条件。

4. 改进建议

除上述缺陷外，测试过程中发现 WIPO Sequence 需要改进的功能。如表 15 所示。

表 15 WIPO Sequence 需要改进的功能

序号	备注/更改请求
1	目前的系统不支持多个申请人/发明人的逐个添加，后添加的申请人/发明人将会替换先添加的申请人/发明人；建议允许添加多个申请人/发明人，可以创建申请人/发明人列表，或者对于多个申请人/发明人的添加规则，系统给予说明。
2	导入多序列格式的符合 ST.25 标准的 txt 文件时，部分氨基酸序列的“三字母”代码未能转换为对应的“单字母”代码；建议系统自动将所有氨基酸序列中的“三字母”转换成对应的“单字母”代码。
3	使用界面的友好程度有待提升，部分按钮的设置较为隐蔽，建议对相应的按钮进行突出显示或差异化设置；部分按钮的翻译不太准确，如“Add Feature”，译为“添加特征”要比“添加功能”更加准确。
4	ST.25 序列中的中文字符（申请人/发明人/发明名称/自由文本）不能被识别，导入后呈现乱码。如 ANSI 编码格式的 ST.25 序列导入后呈现乱码，而 UTF-8 编码格式的 ST.25 序列导入后则能正常显示。 （1）ANSI 编码格式的 ST.25 生物序列： 



（2）UTF-8 编码格式的 ST.25 生物序列：

5. 测试结论

经过测试发现，WIPO Sequence 申请人编著和验证工具能够实现申请端所需要的序列表创建、导入、编辑、验证、导出等功能，支持符合 ST.25 标准的 txt 文件生成 ST.26 标准的 XML 文件，并且能够以人类可读的格式显示生成的符合 ST.26 标准的序列表。

- WIPO Sequence 支持序列表的创建，用户对序列表的通用信息（申请信息、优先权信息、发明名称等）和序列信息（序列名称、分子类型、残基、生物体名称、限定符分子类型等）进行设置后即可创建一个新的序列表。
- WIPO Sequence 支持原始格式、ST.25 格式和 ST.26 格式的序列的导入，成功导入序列后，会出现导入成功的提示框。对于导入的多个序列，可以进行序列的插入、重排、编辑等操作。
- 编辑过程中，对于有特定格式要求的元素/属性，会予以提示；对于有特定取值的元素/属性（如分子序列的取值只能是 DNA、RNA 或 AA）进行了限制，允许申请人通过下拉列表进行选择；对特征关键词及强制性限定词进行了关联设置。
- 对于验证通过（不存在错误提示）的序列表文件，会出现验证通过的提示框，可以生成符合 ST.26 标准的 XML 文件，并能够以人类

可读的格式显示序列。对于验证不通过的序列表文件，则不能生成符合 ST.26 标准的 XML 文件，也不能显示序列；对于序列标准存在的警告或错误以列表的形式进行提示，并附有指定到警告或错误所在位置的链接，方便对序列进行修改。

- 对于创建、导入的序列表文件和生物体文件，用户可以导出。

整体而言，ST.26 的 WIPO Sequence 的功能相对全面，基本能够满足申请人对生物序列的编著和验证需求。但是，目前 WIPO Sequence 是独立的应用程序，如何在智能化升级系统的电子申请、智能审查和智能检索系统中集成，实现相应的导入、编辑、比对和展示等功能，还需要技术人员进一步开发。

(三) WIPO Sequence Validator 的测试与评估

1. 测试环境

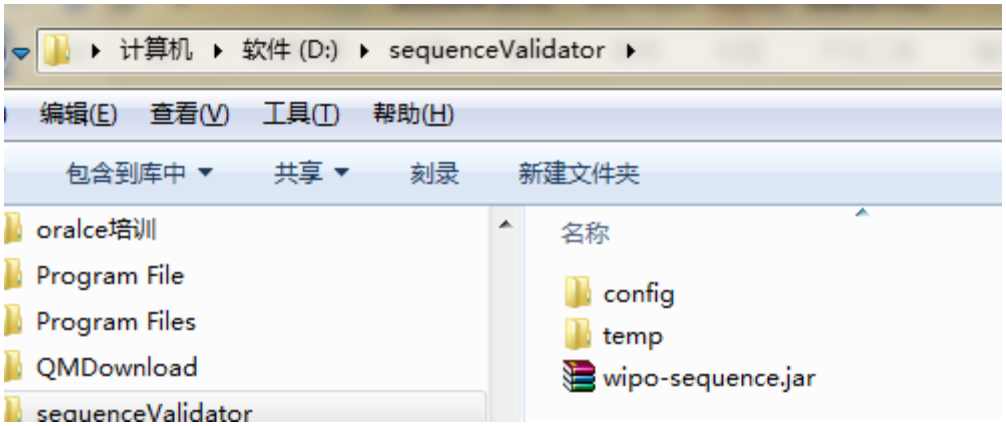
操作系统	浏览器	Java 运行环境	中间件
Windows	chrome	jdk1.8.0_121	Tomcat8.5

2. 测试步骤

本次测试采用两种方式，一种是基于浏览器测试；一种是基于接口 api 调用，测试步骤如下：

(1) 部署 WIPO Sequence Validator 应用

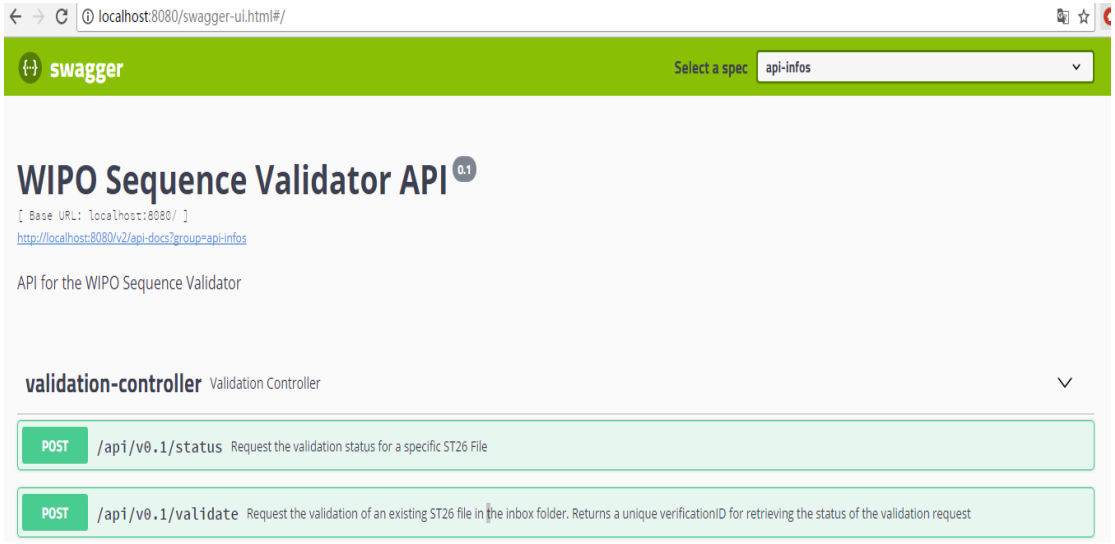
- 拷贝 jar 至某一路径，例如 D:\sequenceValidator。
- 解压 jar 包，获取 application.properties 配置文件，并进行修改。
- 在同一路径下创建 config 文件夹，将修改后的 application.properties 文件放入该文件夹。
- 根据配置 application.properties 信息，创建文件存储文件夹 temp，temp 文件夹包括 inbox、outbox、process 和 report 四个子文件夹。
- 最终目录结构如下：



- 命令行启动应用，java –jar wipo-sequence.jar，启动应用，如下图所示：

(2) 基于浏览器测试

访问 url 为 http://localhost:8080/swagger-ui.html，针对测试用例，进行测试，如下图所示：



(3) 基于接口 api 测试

编写接口调用代码，通过 junit 单元测试进行接口测试，部分测试代码

如下：

```
@Test
public void contextLoads() throws InterruptedException {
    long start = System.currentTimeMillis();
    String id = sequenceValidService2.valid("validator_test01_my_small_not_well_formed.xml", "formality");
    System.out.println("--------------id is ----------------------"+id);
    String rtn = sequenceValidService2.findStatus(id);
    System.out.println("--------------status is ----------------------"+rtn);

    while(true){
        Thread.sleep(2000);
        System.out.println("sssssssssssss");
        System.out.println("--------------status is ----------------------" + sequenceValidService2.findStatus(id));
    }
}
```

3. 测试用例

输入文件	大小	描述
validator_test01_my_small.xml	5k	EPO 小文件
validator_test02_my_large.xml	128k	EPO 大文件
validator_test03_long import.xml	68k	EPO 大文件
validator_test04_huge.xml	5.4M	EPO 超大文件
validator_test01_my_small_data_error.xml	6k	EPO 数据错误文件
validator_test01_my_small_not_well_formed.xml	6k	EPO 格式错误文件
s1-v1.xml	5k	国知局 小文件
s3.xml	4k	国知局 小文件
s6.xml	26k	国知局 大文件

4. 测试结果

基于浏览器测试和基于接口调用 api 测试结果相同，测试类型包括 Full（格式和内容校验）和 Formality（格式校验），测试结果如表 16 和表 17 所示，测试结果文件详见附件“输入输出文件.rar”。

表 16 格式和内容校验结果

输入文件	大小	Valid 接口 Full check	Verification ID	Status 接口 result	Report 文件	结果描述
validator_test01_my_small.xml	5k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING in the report
validator_test02_my_large.xml	128k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING and error in the report
validator_test03_long import.xml	68k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING in the report

validator_test04_huge.xml	5.4M	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING in the report
validator_test01_my_small_data_error.xml	6k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING and error in the report
validator_test01_my_small_not_well_formed.xml	6k	http 400	Not Created		Not Created	File stays in ‘inbox’ folder
s1-v1.xml	5k	202 Accepted	Created	200 FINISHED-VALID	Created	
s3.xml	4k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING in the report
s6.xml	26k	202 Accepted	Created	200 FINISHED-VALID	Created	WARNING and error in the report

表 17 格式校验结果

文件	大小	**Valid 接口 Full check**	**Verification ID**	**Status 接口 result**	**Report 文件**	结果描述
validator_test01_my_small.xml	5k	202 Accepted	Created	200 FINISHED-VALID	Created	
validator_test02_my_large.xml	128k	202 Accepted	Created	200 FINISHED-VALID	Created	
validator_test03_long import.xml	68k	202 Accepted	Created	200 FINISHED-VALID	Created	
validator_test04_huge.xml	5.4M	202 Accepted	Created	200 FINISHED-VALID	Created	
validator_test01_my_small_data_error.xml	6k	202 Accepted	Created	200 FINISHED-VALID	Created	
validator_test01_my_small_not_well_formed.xml	6k	http 400	Not Created		Not Created	File stays in ‘inbox’ folder
s1-v1.xml	5k	202 Accepted	Created	200 FINISHED-VALID	Created	

s3.xml	4k	202 Accepted	Created	200 FINISHED-VALID	Created	
s6.xml	26k	202 Accepted	Created	200 FINISHED-VALID	Created	

5. 测试结论

经过测试发现：WIPO Sequence Validator 目前仅支持符合 ST.26 标准的 XML 文件格式的生物序列表的验证，验证界面不进行错误提示，也不支持在线编辑。如果测试的 XML 文件不符合 ST.26 标准要求的格式，则验证不通过，测试文件仍保留在 inbox 文件夹，也不生成报告；如果测试的 XML 文件符合 ST.26 标准要求的格式，则验证通过，测试文件转移至 outbox 文件夹，对于内容方面存在错误或警告仅以报告（XML 格式）的形式在 report 文件夹中给出。

WIPO Sequence Validator 可以作为微服务部署在服务器环境中，能够与知识产权局使用的其他业务解决方案的应用程序通信，可以对申请人提交序列表数据提供验证服务。但是，WIPO Sequence Validator 的验证报告仅 XML 格式的文件在 report 文件夹中给出，由于 XML 格式的报告不利于人工阅读，因此在智能化升级系统中集成 WIPO Sequence Validator 还需要考虑如何对 XML 格式的文件进行处理，以利于人工阅读的形式进行呈现。

(四) 工具的选择与利弊分析

1. 使用 WIPO 软件工具的利弊

国际局提供免费的 WIPO Sequence 和 WIPO Sequence Validator 工具、以及相应工具更新，但是目前的 WIPO Sequence 是独立的应用程序，如何在智能化升级系统的电子申请、智能审查和智能检索系统中集成，实现序列表的导入、编辑、展示和比对等功能，还需要技术人员进一步开发。WIPO Sequence Validator 的验证报告仅 XML 格式的文件在 report 文件夹中给出。由于 XML 格式的文件不利于人工阅读，因此在智能化升级系统中集成 WIPO Sequence Validator 还需要考虑如何对 XML 格式的文件进行处理，以利于人工阅读的形式进行呈现。

(1) 使用 WIPO 软件工具的优势

国际局提供的 WIPO Sequence 和 WIPO Sequence Validator 工具自 2017 年启动研发，有专门的开发团队，并经历独立团队和终端用户的多轮测试。因此，使用 WIPO 提供的软件工具的优势在于：1）ST.26 相关工具涵盖的功能比较全面；2）ST.26 相关工具是免费的，且提供后续的更新，可以节约成本；3）ST.26 相关工具比较权威，能够确保所提交的序列符合 ST.26 标准。

(2) 使用 WIPO 软件工具的弊端

首先，在智能化升级系统设计中需要考虑集成 WIPO Sequence 和 WIPO Sequence Validator 工具的消耗的人力成本及为满足因标准变化对申请、审查和检索系统提出的新需求而进行的软件开发成本。

其次，使用 WIPO 提供的软件工具将面临技术完全依赖于欧、美等国家，丧失自主研发能力，错失发展的主动权。由于通过知识产权组织的网络获取工具的更新，信息安全方面可能存在隐患。

2. 自主开发软件工具的利弊

(1) 自主开发软件工具的优势

自主开发软件工具的优势主要有两点：一是可以掌握技术的主动权，二是能够保障我国知识产权信息的安全。

(2) 自主开发软件工具的弊端

与 ST.25 标准相比，ST.26 标准在适用范围、文件格式和标准内容等方面的变化较大。这些变化及相应的功能需要在自主开发的软件工具中全面实现，不是一朝一夕的事情。例如，WIPO 的 ST.26 工具的开发耗时 3 年。因此，自主开发软件工具的弊端主要是开发软件工具需要耗费较大的人力、物力、财力和时间成本。

第六章 总结与展望

一、 本研究的主要结论

本研究通过文献研究、调查研究、比较研究和专家论证等方法的综合运用，在对 ST.26 标准和 ST.25 标准进行全面分析和比较的基础上，考察 ST.26 标准相对于 ST.25 标准的变化，以及这些变化对于智能化升级系统设计的影响。本研究的主要结论有：

(一) ST.26 标准简述

ST.26 标准源于 2010 年欧洲专利局（EPO）提出的建立一个基于 XML 格式的生物序列标准，其目的在于使申请人在专利申请中所撰写的序列表能在国际和国家阶段都被接受；提高序列表述的精确度和质量，从而有利于申请人、公众和审查员更容易传播该序列；有利于序列数据的检索；以及允许序列数据以电子形式进行交换和输入计算机数据库。

符合 ST.26 标准的序列表包括通用数据部分和序列数据部分。通用部分主要包括著录信息，用于与该序列表对应的专利申请进行关联；序列数据部分主要由关于该序列信息的一个或多个数据元素组成，数据元素又包括各种特征关键词和限定词。

(二) ST.26 相对于 ST.25 标准的主要变化

与 ST.25 标准相比，ST.26 标准的变化主要体现在适用范围、数据格式、标准内容等方面。

从适用范围来看，ST.26 标准的适用范围更加广泛，不仅适用于国际申请，也适用于国内或区域程序的专利申请。

从数据格式来看，ST.26 标准采用了 XML 格式，在数据管理、交换和解析方面更加安全、便捷。

从标准内容来看，ST.26 标准的著录信息和序列信息更加丰富，采用特征关键词和特征限定词共同描述序列的特征，序列的表示更加详细和专业。

(三) 标准变化对智能化升级系统设计的影响

ST.26 标准与 ST.25 标准之间的差异对系统设计的影响，主要体现在应用系统设计、数据库设计和历史数据处理三个方面。在电子申请系统的设计中，需要在客户端及交互式平台中提供符合 ST.26 标准的序列表的生成、导入、编辑和校验功能；在智能审查系统中，需要提供符合 ST.26 标准的序列表的校验、形式审查及展示功能；在智能检索系统中，需要提供代码转换、序列比对和序列翻译功能。在数据库设计方面，需要包含 ST.26 标准的 DTD 规范中的重要数据项，在数据字段的设置、字段长度、属性及其取值与 ST.26 标准的要求相符合。此外，在智能化系统升级的设计中还要考虑历史数据的处理问题。

(四) 中国生物序列的现状及未来

当前，中国的生物序列主要按照 ST.25 标准执行，使用过程中也存在诸多问题，例如 CEPCT 电子申请客户端不支持生物序列的编辑和校验，CPC 电子申请客户端的序列表验证程序偶尔出现错误，PCT 国际申请中序列表因格式和要求不一致而只能以光盘提交，检索系统缺少序列转换、序列比对和序列翻译等功能。虽然，知识产权出版社曾基于 ST.26 标准草案开发了符合 ST.26 标准的数据加工规则、加工方式，但其开发的软件工具目前仅限于后端的加工环节，而不涉及申请端、审查端和检索端，其能否应用于智能化升级系统还需要进行深入地研判。

根据国际局提供的 ST.25 向 ST.26 过渡的路线图，中国的生物序列在不久的将来，必然要符合 ST.26 标准。因此，在 2022 年 1 月 1 日来临之前，中国需要从计划制定、规则修改、工具测试、IT 系统升级和人员培训等方面做好准备。

(五) ST.26 标准配套软件工具的评估

ST.26 标准配套的软件工具由 WIPO Sequence（申请人编著和验证程序），WIPO Sequence Validator（局端验证程序），以及 WIPO Sequence Server（更新和发布服务器）组成。

WIPO Sequence 能够实现申请端所需的序列创建、导入、编辑、验证、导出等功能，支持符合 ST.25 标准的 txt 文件生成 ST.26 标准的 XML 文件，基本能够满足申请人对生物序列的编著和验证需求。但是，目前的 WIPO Sequence 是独立的应用程序，如何在智能化升级系统的电子申请、智能审查和智能检索系统中集成，实现相应的导入、编辑、展示等功能，还需要技术人员进一步开发。

WIPO Sequence Validator 可以作为微服务部署在服务器环境中，与知识产权局使用的其他业务解决方案的应用程序通信，可以对申请人提交序列表数据提供验证服务。但是，WIPO Sequence Validator 目前仅支持符合 ST.26 标准的 XML 文件格式的生物序列表的验证，验证报告仅 XML 格式的文件在 report 文件夹中给出。由于 XML 格式不利于人工阅读，因此在智能化升级系统中集成 WIPO Sequence Validator，还需要考虑如何对 XML 格式的文件进行处理，以利于人工阅读的形式进行呈现。

二、 本研究的不足和展望

受研究资料和工具的限制，本研究主要存在以下不足：

首先，本研究中考虑标准变化对智能化升级系统设计的影响是基于已有资料，从应用系统的申请端、审查端和检索端需要具备的功能模块进行的；生物序列数据库建设和历史数据处理也只是从符合 ST.26 标准要求的角度提出了原则性建议。

其次，国际局开发的 ST.26 标准的配套软件工具尚未正式发布，本研究中对 ST.26 配套软件工具的测试和评估是基于 v0.9.9 测试版本进行的。待 ST.26 标准的配套软件工具正式发布后，还需要进行系统地测试和评估，从功能实现的角度评估智能化升级系统中集成和改造的可行性。

再次，对于中国生物序列数据库的建设，不仅要在字段及其属性的设置上包含 ST.26 标准给出的 DTD 规范中的重要数据项，满足 ST.26 标准的要求，还要考虑如何将中国生物序列数据与专利数据、国际通用生物序列数据等进行关联，以便于检索和利用。

参考文献

[1] Osmat A. Jefferson, Deniz Köllhofer, Prabha Ajjikuttira, et al. Public disclosure of biological sequences in global patent practice[J], World Patent Information, 2015 (43):12-24.
[2] Poliana Belisário Zorzal, Fabricia Pires Pimenta,Antonio Alberto Ribeiro Fernandes, et al. Sufficiency of disclosure and genus claims for protection of biological sequences: a comparative study among the patent offices in Brazil, Europe and the United States[J]. Biotechnology Research and Innovation, 2019,3(1): 91-102.
[3] WIPO 标准使用情况的调查[EB/OL]. [2019-06-27]. https://www3.wipo.int/confluence/display/usestandards/WIPO+Standard+ST.25%3A+Presentation+of+nucleotide+and+amino+acid+sequence+listings.
[4] 关于知识产权行业标准《电子文件标准》中部分特征关键词表的修订（第 248 号）. [2019-07-17].http://www.sipo.gov.cn/docs/pub/old/zwgg/gg/201707/t20170726_1312897.html.
[5] 美国《联邦行政法典》中有关核苷酸序列和/或氨基酸序列的规定[EB/OL]. [2019-08-08].https://www.uspto.gov/web/offices/pac/mpep/mpep-9020-appx-r.html#d0e333653.
[6] 欧洲专利授权程序中关于序列表的规定[EB/OL]. [2019-06-27]. http://archive.epo.org/epo/pubs/oj013/11_13/11_5423.pdf.
[7] 曲超, 张松, 曲晓光. WIPO 标准 ST.26 与 ST.25 对比研究分析[J]. 中国标准化, 2012(9):50-52.
[8] 张春华, 马晓蕾, 李茉. WIPO 标准 ST.25 与 ST.26 差异分析[J]. 标准科学, 2015(2):67-71.
[9] 中国专利生物序列数据标引技术服务合同(2017-2018)[EB/OL]. [2019-06-20] http://www.sipo.gov.cn/ztzl/zfcgzl/htgs/1124874.htm.
[10] 《专利审查指南 2010》[EB/OL].[2019-07-17]. http://www.sipo.gov.cn/zhfwpt/zlsqzn_pt/zlfssxzjsczn/index.htm.

致　谢

本研究是中国专利信息中心承接的局智能化升级审查小组提出的研究任务。局智能化升级审查小组的信任给予了研究人员莫大的精神鼓舞。

信息中心承接该任务后，中心主任刘燕新、加工部部长方建国和信息研究处处长杨筱高度重视，立即展开部署，组织研究人员积极投入工作。在本研究工作即将完成之际，特别感谢他们为本研究工作提供的便利条件和悉心指导。

在研究过程中，国家知识产权局自动化部对外交流处提供了丰富的研究资料，为本研究工作提供了基础支撑，在此特别提出感谢！

感谢在调研过程中提供支持的许家升、李进两位同志，以及被调查的代理所和国知局相关人员的积极配合和热心帮助。

感谢在工具测试过程中提供技术支持的商小奇和路小霞两位同事！

感谢研究组成员费一楠、扈林芳、王天鹤、熊熙然、田欣等的鼎力支持，正是他们的辛勤付出和协作保证了课题研究的顺利进行！