ICS 35.240.80
C 07

团　　体　　标　　准

T/CHIA 42.3-2023

长非编码 RNA 和蛋白相互作用注释标准 第 3 部分：长非编码 RNA 及其相互作用蛋白质的功能注释

Specifications for interaction between long non-coding RNA and proteins

Part 3：Functional annotation of long non-coding RNA and proteins

2023-11-14 发布　　2024-02-01 实施

中国卫生信息与健康医疗大数据学会　发布

目　次

前　言

本标准按照GB/T 1.1—2020给出的规则起草。

T/CHIA 42-2023《长非编码RNA和蛋白相互作用注释标准》分为以下3个部分：

——第1部分：RIP-seq和CLIP-seq的实验方法流程

——第2部分：RIP-seq和CLIP-seq的数据分析方法

——第3部分：长非编码RNA及其相互作用蛋白质的功能注释

本标准为T/CHIA 42-2023的第3部分。

本标准由中国科学院生物物理研究所提出，由中国卫生信息与健康大数据学会归口。

本标准主要起草单位：中国科学院生物物理研究所、中国科学院北京基因组研究所（国家生物信息中心）、浙江大学、复旦大学、清华大学、中国人民解放军总医院、北京蛋白质组研究中心、中国科学院微生物研究所、北京大学人民医院、中国科学院上海营养与健康研究所、中南大学、空军军医大学（第四军医大学）、中国科学院计算技术研究所和北京睿博解码生物科技有限公司。

本标准主要起草人：陈润生、何顺民、宋廷瑞、张鹏、周红红、王晓娜、方向东、金力、何昆仑、李亦学、张学工、段会龙、周水庚、渠鸿竹、赵思琪、钱颖、王霞、赵屹、吕旭东、朱云平、马俊才、杨忠、石乐明、吴松峰、吴林寰、王振、陈先来、贾志龙、张昭军、娄晓敏、阮修艳、单广乐、乔楠、刘登辉、丁子建。

引　言

《长非编码 RNA 与蛋白相互作用注释标准 第 3 部分：长非编码 RNA 及其相互作用蛋白质的功能注释》旨在规范现有长非编码 RNA 和蛋白质相互作用的功能注释，为长非编码 RNA 和蛋白质相互作用的注释信息提供一套术语规范、定义明确、语义语境无歧义的标准。

长非编码 RNA 和蛋白相互作用注释标准 第 3 部分：长非编码 RNA 及其相互作用蛋白质的功能注释

1 范围

本标准规定了长非编码 RNA 及蛋白质相互作用的功能注释规范。

本标准适用于指导研究人员对长非编码 RNA 及蛋白质相互作用进行注释和整理。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本标准必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本标准；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

GB/T 29859-2013 生物信息学术语

GB/T 30989-2013 高通量基因测序技术规程

GB/T 35890-2018 高通量测序数据序列格式规范

3 术语和缩略语

GB/T 29859-2013 中界定的术语和定义适用于本标准。

4 长非编码 RNA 与蛋白质相互作用注释元数据

4.1 元数据描述

本部分采用摘要表示的方式定义和描述元数据。摘要内容包括以下 7 个属性：中文名称、定义、英文名称、数据类型、值域、约束/条件、备注。

4.1.1 中文名称

指元数据元素或元数据实体的中文名称。

4.1.2 定义

描述元数据元素或元数据实体的基本内容，给出信息资源某个特性的概念和说明。

4.1.3 英文名称

元数据元素或元数据实体的英文名称，一般用英文全称

4.1.4 数据类型

元数据元素的数据类型，对元数据元素的有效值域和允许对该值域内的值进行有效操作的规定。

4.1.5 值域

说明元数据元素可以取值的范围。

4.1.6 约束/条件

说明元数据实体或元数据元素是否必须选取的属性。包括必选、可选。

必选（M）：表明该元数据元素或元数据实体必须选择。

可选（O）：根据实际应用可以选择也可以不选的元数据元素或元数据实体。如果一个可选元数据实体未被使用，则该实体所包含的元素（包括必选元素）也不选用。可选元数据实体可以有必选元素，但只当可选实体被选用时才成为必选。

4.1.7 备注

对元数据元素或元数据实体的进一步说明（根据需要选用）

4.2 元数据

本节给出了长非编码 RNA 与蛋白质相互作用元数据的定义。本标准三个元数据之间的关系如图 1 所示：

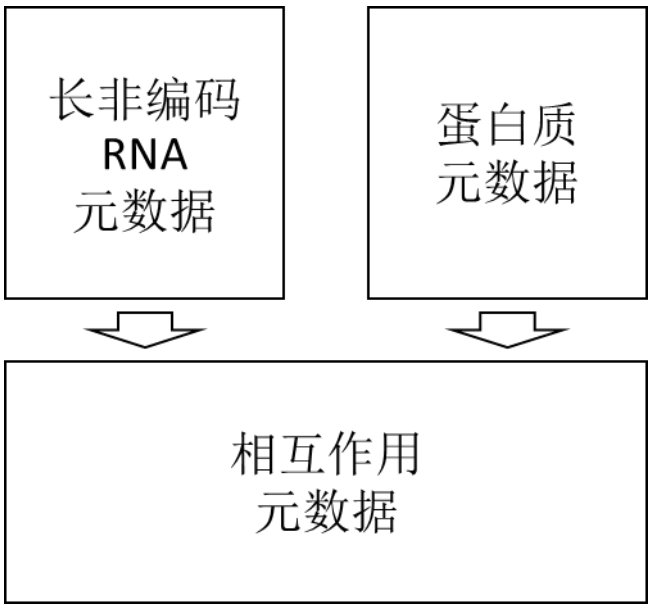

图 1 元数据的层级关系

具体内容可参考 NPInter 数据库中的实例，如其中长非编码 RNA MALAT1 与蛋白 AN KHD1 的相互作用相关数据内容在如下链接中查看： http://bigdata.ibp.ac.cn/npinter5/interaction/ncRI-51495421/。

4.2.1 相互作用长非编码 RNA

相互作用的长非编码 RNA 的注释信息包括长非编码 RNA 的名称、数据库编号、收录数据库、物种、描述以及相关功能注释。

4.2.1.1 名称

定义：长非编码 RNA 的标准名称

英文名称：gene symbol

数据类型：字符串

值域：数据库中的长非编码 RNA 的名称

约束条件：必选项（M）

4.2.1.2 收录数据库

定义：收录该长非编码 RNA 的数据库
英文名称：database
数据类型：字符串
值域：数据库名称
约束条件：必选项（M）
备注：推荐使用 Ensembl 数据库和 NONCODE 数据库

4.2.1.3 数据库编号

定义：长非编码 RNA 在收录数据库里面的唯一标识
英文名称：database ID
数据类型：字符串
值域：数据库中长非编码 RNA 的唯一标识
约束条件：必选项（M）

4.2.1.4 物种名

定义：长非编码 RNA 所在的物种名称
英文名称：organism
数据类型：字符串
值域：物种标准拉丁名称
约束条件：必选项（M）

4.2.1.5 描述

定义：长非编码 RNA 的描述
英文名称：description
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.1.6 功能注释

定义：长非编码 RNA 的功能
英文名称：function annotation
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.1.7 疾病注释

定义：长非编码 RNA 相关的疾病信息
英文名称：disease annotation
数据类型：字符串
值域：自由文本
约束条件：可选项（O）
备注：疾病名称从 human disease ontology（DOID）中选择

4.2.2 相互作用蛋白质

相互作用的蛋白质的注释信息包括蛋白质的名称、数据库编号、物种、描述以及相关功能注释。

4.2.2.1 名称

定义：蛋白质所在基因的标准名称
英文名称：gene symbol
数据类型：字符串
值域：基因标准名称
约束条件：必选项（M）

4.2.2.2 数据库编号

定义：蛋白质在 Uniprot 数据库里面的唯一标识
英文名称：database ID
数据类型：字符串
值域：数据库中蛋白的唯一标识
约束条件：必选项（M）
备注：选用 UniProt 数据库的 accession 信息作为数据库编号

4.2.2.3 物种名

定义：蛋白质所在的物种名称
英文名称：organism
数据类型：字符串
值域：物种标准拉丁名称
约束条件：必选项（M）

4.2.2.4 描述

定义：蛋白质的描述
英文名称：description
数据类型：字符串
值域：自由文本
约束条件：必选项（M）
备注：必须包括蛋白质的全名

4.2.2.5 功能注释

定义：蛋白质的功能
英文名称：function annotation
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.2.6 疾病注释

定义：蛋白质相关的疾病信息
英文名称：disease annotation

数据类型：字符串
值域：自由文本
约束条件：可选项（O）
备注：疾病名称从 human disease ontology（DOID）中选择

4.2.3 相互作用

相互作用的注释信息包括注释标识、物种、作用分子、发现组织/细胞系、来源、描述、参考文献等。

4.2.3.1 注释标识

定义：相互作用的唯一标识符
英文名称：interaction ID
数据类型：字符串
值域：自定义或引用数据库中的相互作用标识符
约束条件：必选项（M）

4.2.3.2 物种名

定义：发现该相互作用所在的物种名称
英文名称：organism
数据类型：字符串
值域：物种标准拉丁名称
约束条件：必选项（M）

4.2.3.3 相互作用蛋白名称

定义：相互作用蛋白的所在基因名称
英文名称：interacted protein gene symbol
数据类型：字符串
值域：相互作用蛋白质元数据中的名称
约束条件：必选项（M）

4.2.3.4 相互作用长非编码 RNA 名称

定义：相互作用长非编码 RNA 的所在基因名称
英文名称：interacted lncRNA gene symbol
数据类型：字符串
值域：相互作用长非编码 RNA 元数据中的名称
约束条件：必选项（M）

4.2.3.5 组织/细胞系

定义：发现该相互作用所在的细胞系或者组织名称
英文名称：tissue/cell line
数据类型：字符串
值域：自由文本
约束条件：必选项（M）
备注：如果使用细胞系，需要使用标准的细胞系名称

4.2.3.6 来源

定义：获取该相互作用的来源
英文名称：data source
数据类型：字符串
值域：自由文本
约束条件：必选项（M）

4.2.3.7 描述

定义：对该相互作用的描述信息
英文名称：description
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.3.8 参考文献

定义：该相互作用的来源文献
英文名称：reference
数据类型：字符串
值域：自由文本
约束条件：必选项（M）
备注：可以有多条出版物信息

4.2.3.8.1 PubMed 序列号

定义：文章在 PubMed 库的编号
英文名称：PubMed ID
数据类型：字符串
值域：PubMed 文献唯一标识符
约束条件：可选项（O）

4.2.3.8.2 DOI 编号

定义：文章的 DOI 编号
英文名称：DOI
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.3.8.3 期刊名称

定义：文章发表的期刊名称
英文名称：journal name
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.3.8.4 文章标题

定义：文章标题
英文名称：article title
数据类型：字符串
值域：自由文本
约束条件：可选项（O）

4.2.3.8.5 发表年份

定义：文章发表的年份
英文名称：year of publication
数据类型：字符串
值域：自由文本
约束条件：可选项（O）
备注：格式为“yyyy”，如“2018”

4.2.3.9 长非编码 RNA 结合位点

定义：相互作用中，RNA 与蛋白质结合的位点
英文名称：lncRNA binding site
数据类型：字符串
值域：格式为“x-x”，x 为数字，代表从 5’端开始的核苷酸位置信息，如“1-200”
约束条件：可选项（O）

4.2.3.10 蛋白质结合位点

定义：相互作用中，蛋白质与 RNA 结合的位点
英文名称：protein binding site
数据类型：字符串
值域：格式为“x-x”，x 为数字，代表从 5’端开始的氨基酸位置信息，如“1-200”
约束条件：可选项（O）