ICS 35.240.80
C 07

# 团 体 标 准

T/CHIA 41.3-2023

# 非编码 RNA 注释标准
# 第 3 部分：功能及疾病注释

Specifications for annotation of non-coding RNA
Part 3：The function and diseases

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

T/CHIA 41-2023《非编码RNA注释标准》分为以下3个部分：

——第1部分：分类信息；

——第2部分：基本信息注释；

——第3部分：功能及疾病注释。

本部分为T/CHIA 41-2023的第3部分。

本部分按照 GB/T 1.1-2020给出的规则起草。

本部分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。

本部分主要起草单位：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）和北京睿博解码生物科技有限公司。

本部分的主要起草人：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、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

《非编码 RNA 注释标准 第 3 部分：功能及疾病注释》旨在为非编码 RNA 的注释提供一套术语规范、条理清晰、意义明确、语义语境无歧义的关于非编码功能及与疾病关系的注释标准，防止注释者在对非编码 RNA 进行功能注释过程中遗漏相关信息，并保证注释信息的一致性，便于对非编码 RNA 已知知识的获取，保证信息的有效交换、有效挖掘分析和共享。

# 非编码 RNA 注释标准
# 第 3 部分：功能及疾病注释

## 1 范围

本标准规定了非编码RNA注释中关于功能部分以及与疾病关系的注释。

本标准适用于对非编码 RNA 注释。

## 2 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准的条款。凡是注日期的引用文件，其随后所有的修改单（不包括勘误的内容）或修订版均不适用于本标准。凡是不注日期的引用文件，其最新版本适用于本标准。

GB/T 14396-2016 疾病分类与代码

GB/T 29859-2013 生物信息学术语

T/CHIA 21.5-2021 组学样本处理与数据分析标准 第 5 部分：转录组测序数据分析

## 3 术语和定义

GB/T 29859-2013、T/CHIA 21.5-2021 中界定的以及下列术语和定义适用于本文件。

3.1

**保守性** conservation

是进化上的一个概念，指在生物进化的过程中，某些生物大分子或细胞结构等基本没有变化或变化不明显，较稳定地存在。

3.2

**表达谱芯片** RNA microarray

采用 cDNA 或寡核苷酸片段作为探针，用核酸探针杂交的原理来检测表达水平的变化。

3.3

**基因本体论** Gene ontology

有关基因功能描述的知识数据库，详情见附录 6.1。

## 4 缩略语

下列缩略语适用于本标准。

cDNA：互补脱氧核糖核酸（complementary DNA）。

FPKM：每千碱基外显子长度、每百万总比对片段的转录本片段数（Fragments Per

Kilobase of exon model per Million mapped fragments）。

TPM：每百万总转录本数量的单个转录本数量（Transcripts Per Million）。

CPM：每百万总比对片段的片段个数（Counts Per Million）。

GO：基因本体论（Gene ontology）。

BP：生物过程（Biological Process）。

CC：细胞组分（Cellular Component）。

MF：分子功能（Molecular Function）。

KEGG：生物代谢通路数据库（Kyoto Encylopedia of Genes and Genomes）。

## 5 非编码 RNA 功能及疾病注释元数据

### 5.1 元数据描述

与本标准第二部分6.1一致，摘要内容包括7个属性：中文名称、定义、英文名称、数据类型、约束/条件、词表、备注。

### 5.2 元数据

本节给出非编码 RNA 功能及疾病注释相关元数据的定义。

#### 5.2.1 功能及疾病注释

功能及疾病注释元数据提供了本标准功能及疾病注释的总体框架，包括基因识别号、表达谱、共表达基因、相互作用、靶基因、保守性、基因本体论、代谢通路、编码潜力、变异列表以及疾病关系等。

表1. 功能及疾病注释元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 本注释针对的基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 表达谱 | 非编码RNA在各组织样本中的表达量 | Expression profile | 元数据列表 | 可选（O） | 依据本标准的表达量元数据定义 | 可列举多个 |
| 共表达基因 | 非编码RNA与其它基因的共表达关系 | Co-expression | 元数据列表 | 可选（O） | 依据本标准的共表达基因元数据定义 | 可列举多个 |
| 相互作用 | 非编码RNA与其它分子的相互作用 | Interaction | 元数据列表 | 可选（O） | 依据本标准的相互作用元数据定义 | 可列举多个 |
| 靶基因 | 受非编码RNA调控影响的基因 | Target gene | 元数据列表 | 可选（O） | 依据本标准的靶基因元数据定义 | 可列举多个 |

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 保守性 | 非编码RNA在不同物种间的保守程度 | Conservation | 元数据 | 可选（O） | 依据本标准的保守性元数据定义 | |
| 基因本体论 | 非编码RNA相关基因的功能描述 | GO | 字符串 | 可选（O） | GO数据库的标识符，参考附录6.1，如：GO:0001591 | 可列举多个 |
| 代谢通路 | 非编码RNA参与的KEGG代谢通路 | KEGG pathway | 字符串 | 可选（O） | KEGG数据库的通路标识符，如：hsa00010 | 可列举多个 |
| 编码潜力 | 非编码RNA潜在编码蛋白的能力 | Coding potential | 字符串 | 可选（O） | 编码潜力评分，翻译活动证据，潜在编码区域等 | |
| 变异列表 | 非编码RNA上与参考基因组不一致的位点 | Variants | 元数据列表 | 可选（O） | 依据本标准的变异位点元数据定义 | 可列举多个 |
| 疾病关系 | 非编码RNA与疾病的关系 | Relationship to disease | 元数据列表 | 可选（O） | 依据本标准的疾病关系元数据定义 | 可列举多个 |

### 5.2.2 表达量

表达量元数据记录非编码RNA在特定组织样本中的表达水平，包括基因识别号、转录本识别号、样本类型、数据类型、表达量单位、表达值以及数据来源。

表2. 表达量元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 表达量对应的基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 转录本识别号 | 表达量对应的转录本 | Transcript identity | 字符串 | 可选（O） | 本标准第二部分转录本元数据中的转录本识别号 | 用于转录本水平的表达量 |
| 样本类型 | 表达量来源的样本类型 | Sample type | 字符串 | 必选（M） | 如：血液，心脏，某细胞系等 | |
| 数据类型 | 检测表达量的技术和数据形式 | Data type | 字符串 | 必选（M） | 如：转录组测序，表达谱芯片，PCR等，可包含技术细节，如总RNA，双末端等 | |
| 表达量单位 | 表达值的单位，反映数据处理方式 | Expression unit | 字符串 | 必选（M） | 如：RPKM，FPKM，CPM，count，log2亮度值，CT值等 | |
| 表达值 | 表达量的具体数值 | Expression value | 数字 | 必选（M） | 依据表达量单位的不同，可能为整数或实数，有时可为负值 | |

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 数据来源 | 表达谱数据的来源 | Data source | 字符串 | 可选（O） | 如数据库收录号 NCBI: GSM6476844 | |

### 5.2.3 共表达基因

共表达基因元数据记录非编码RNA与其它基因的共表达关系，包含基因识别号、转录本识别号、共表达基因识别号、共表达基因名称、共表达转录本、样本类型、相关性、显著性以及数据来源。

表3. 共表达基因元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 非编码RNA基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 转录本识别号 | 非编码RNA转录本 | Transcript identity | 字符串 | 可选（O） | 本标准第二部分转录本元数据中的转录本识别号 | 用于转录本水平的共表达 |
| 共表达基因识别号 | 与非编码RNA共表达的基因识别号 | Co-expressed gene identity | 字符串 | 必选（M） | 编码或非编码基因的识别号 | |
| 共表达基因名称 | 与非编码RNA共表达的基因名称 | Co-expressed gene symbol | 字符串 | 可选（O） | 编码或非编码基因的名称，参考HGNC命名 | |
| 共表达转录本 | 与非编码RNA共表达的转录本识别号 | Co-expressed transcript | 字符串 | 可选（O） | 编码或非编码转录本的识别号 | 用于转录本水平的共表达 |
| 样本类型 | 分析共表达依据的表达谱数据样本类型 | Sample type | 字符串 | 可选（O） | 如：血液，心脏，某细胞系等 | 可多组织联合分析共表达 |
| 相关性 | 共表达的相关性数值 | Correlation | 数字 | 可选（O） | 相关系数R或类似的统计量 | |
| 显著性 | 共表达的显著性 | Significance | 数字 | 可选（O） | 代表显著性的P值或校正后的FDR | |
| 数据来源 | 分析共表达依据的表达谱数据的来源 | Data source | 字符串 | 可选（O） | 如数据库收录的数据集编号NCBI: GSE211551或论文编号等 | |

### 5.2.4 相互作用

相互作用元数据记录非编码RNA与其它生物分子的相互作用关系，包含基因识别号、转录本识别号、分子类型、分子识别号、分子名称、结合位点、样本类型以及数据来源。

表4. 相互作用元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 非编码RNA基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 转录本识别号 | 非编码RNA转录本 | Transcript identity | 字符串 | 可选（O） | 本标准第二部分转录本元数据中的转录本识别号 | 用于转录本水平的相互作用 |
| 分子类型 | 与非编码RNA相互作用的分子类型 | Molecular type | 字符串 | 必选（M） | 如蛋白，mRNA，miRNA，DNA等 | |
| 分子识别号 | 与非编码RNA相互作用的分子的识别号 | Molecular identity | 字符串 | 可选（O） | 如蛋白识别号，mRNA识别号等 | 无识别号等情况（如DNA）可不填 |
| 分子名称 | 与非编码RNA相互作用的分子的名称 | Molecular name | 字符串 | 可选（O） | 相互作用分子的常用名称 | |
| 结合位点 | 相互作用分子接触结合的关键位点 | Binding sites | 字符串 | 可选（O） | 具体位置或区间，如转录本或蛋白上的位置，DNA结合位点等 | 包括非编码RNA和互作分子的位置 |
| 样本类型 | 相互作用所在的样本类型 | Sample type | 字符串 | 可选（O） | 如：血液，心脏，某细胞系等 | |
| 数据来源 | 分析相互作用的数据的来源 | Data source | 字符串 | 可选（O） | 如数据库收录的数据编号NCBI: GSE28180或论文编号等 | |

### 5.2.5 靶基因

靶基因元数据记录受非编码RNA调控的基因，包含基因识别号、转录本识别号、调控类型、靶基因识别号、靶基因名称、靶基因转录本、调控位点、样本类型以及数据来源。

表5. 靶基因元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 非编码RNA基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因 | |

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| | | | | | 识别号 | |
| 转录本识别号 | 非编码RNA转录本 | Transcript identity | 字符串 | 可选（O） | 本标准第二部分转录本元数据中的转录本识别号 | 用于转录本水平的调控关系 |
| 调控类型 | 非编码RNA调控的机制类型 | Target type | 字符串 | 可选（M） | 如RISC，翻译抑制，海绵，转录增强，表观沉默等等 | |
| 靶基因识别号 | 受非编码RNA调控的靶基因的识别号 | Target gene identity | 字符串 | 必选（O） | 编码基因或非编码RNA基因的识别号 | |
| 靶基因名称 | 受非编码RNA调控的靶基因的名称 | Target gene name | 字符串 | 可选（O） | 靶基因的常用名称 | |
| 靶基因转录本 | 受非编码RNA调控的转录本 | Target transcript | 字符串 | 可选（O） | 靶基因转录本的识别号 | 用于转录本水平的调控关系 |
| 调控位点 | 调控时结合的关键位点 | Target sites | 字符串 | 可选（O） | 具体位置或区间，如miRNA靶位点，DNA结合位点等 | |
| 样本类型 | 调控关系所在的样本类型 | Sample type | 字符串 | 可选（O） | 如：血液，心脏，某细胞系等 | |
| 数据来源 | 调控关系依据的数据来源 | Data source | 字符串 | 可选（O） | 如数据库编号或论文编号等 | |

### 5.2.6 保守性

保守性元数据记录非编码RNA在不同物种间的保守程度以及各物种中对应的基因和位置区间，包括基因识别号、保守性得分、同源基因和同源位置。

表6. 保守性元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 非编码RNA基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 保守性得分 | 非编码RNA基因序列的保守性分值 | Conservation score | 数字 | 可选（O） | 如PhastCons等算法提供的多物种保守性得分均值 | |
| 同源基因 | 其它物种与非编码RNA同源的基因 | Homologous gene | 字符串 | 可选（O） | 格式可为 物种：基因识别号 | 可列举多个 |

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 同源位置 | 其它物种与非编码RNA同源的染色体区间 | Homologous region | 元数据列表 | 可选（O） | 依据本标准第二部分的基因组位置元数据定义，列举其它物种基因组的位置 | 可列举多个 |

### 5.2.7 变异位点

变异位点元数据记录非编码RNA上的变异位点信息，包括基因组版本、染色体、染色体起始位点、变异识别号、参考序列、变异序列、变异频率以及变异注释。

表7. 变异位点元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因组版本 | 基因组位置所依据的参考基因组组装版本 | Assembly | 字符串 | 必选（M） | 参考NCBI物种分类数据库和UCSC基因组浏览器中的参考基因组命名，如GRCh38，hg19，mm10等 | |
| 染色体 | 变异所在的染色体名称 | Chromosome | 字符串 | 必选（M） | 基因组序列fasta文件中的序列名，如chrX，1，NC_000001.11等 | |
| 染色体起始位点 | 变异在染色体上的5’端位置 | Chromosome Start site | 整数 | 必选（M） | 染色体上的最小位置，通常为一基坐标 | |
| 变异识别号 | 变异的唯一标识符 | Variant identity | 字符串 | 可选（O） | 参考NCBI dbSNP数据库命名，如 rs59306077 | |
| 参考序列 | 变异在参考基因组上的对应序列 | Reference sequence | 字符串 | 必选（M） | 参考基因组上变异对应的参考序列 | |
| 变异序列 | 变异后的序列 | Alternative sequence | 字符串 | 必选（M） | 变异后的实际正链序列 | 同位置可有多种变异，逗号分隔 |
| 变异频率 | 人群中变异出现的频率 | Allele frequency | 字符串 | 可选（O） | 如AF=0.123，可列举多个人群的频率 | 多种变异用逗号分隔 |
| 变异注释 | 变异位点的注释信息 | Variant annotation | 字符串 | 可选（O） | 变异的注释信息，如转录本上的位置，对氨基酸编码的影响，已知的疾病关联等 | |

### 5.2.8 疾病关系

疾病关系元数据记录非编码RNA与疾病的联系，包括基因识别号、转录本识别号、疾病名称、疾病分期、关系描述、临床用途、检测指标、样本类型以及数据来源。

表8. 疾病关系元数据

| 中文名称 | 定义 | 英文名称 | 数据类型 | 约束条件 | 词表 | 备注 |
|---|---|---|---|---|---|---|
| 基因识别号 | 非编码RNA基因 | Gene identity | 字符串 | 必选（M） | 本标准第二部分非编码RNA基因元数据中的基因识别号 | |
| 转录本识别号 | 疾病关联的转录本 | Transcript identity | 字符串 | 可选（O） | 本标准第二部分转录本元数据中的转录本识别号 | 用于转录本水平的疾病关系描述 |
| 疾病名称 | 非编码RNA相关的疾病的名称 | Disease name | 字符串 | 必选（M） | 参考国家标准GB/T 14396-2016中的疾病分类与编号 | |
| 疾病分期 | 非编码RNA相关的疾病进展阶段 | Disease stage | 字符串 | 可选（O） | 根据具体疾病的相关分期标准 | |
| 关系描述 | 非编码RNA与疾病关系的具体描述 | Relationship description | 字符串 | 可选（O） | 如某癌症中某RNA高表达，某RNA变异导致某疾病风险增加等等 | |
| 临床用途 | 非编码RNA在临床上的应用场景 | Clinical application | 字符串 | 可选（O） | 如辅助诊断，治疗靶点，用药指导，预后等，可具体描述 | |
| 检测指标 | 临床中的检测标准及对应临床意义 | Test parameter | 字符串 | 可选（O） | 如表达量或多基因评分超过某值则预后差，检测到某变异不宜用某药等 | |
| 样本类型 | 疾病关系所检测的样本类型 | Sample type | 字符串 | 可选（O） | 如：血液，心脏，皮肤等 | |
| 数据来源 | 疾病关系依据的数据来源 | Data source | 字符串 | 可选（O） | 如数据库编号或论文编号等 | |

## 6 附录

### 6.1 基因本体论

#### 6.1.1 概述

基因本体论是有关基因功能描述的知识数据库，其对生物学领域从分子功能、细胞组分和生物过程三个方面进行了描述。在对非编码RNA的功能注释中，一般是通过对非编码

RNA的靶基因或者非编码RNA共表达的基因进行功能注释，进而达到对非编码RNA可能的功能注释。

6.1.2 分子功能

单个或多个基因产物的复合物在分子水平上的活动，比如蛋白激酶（protein kinase）具有GO分子功能：蛋白激酶活性（protein kinase activity）。

6.1.3 细胞组分

基因产物在执行功能时所处的细胞结构位置，比如核糖体、细胞核。细胞组分是细胞解剖结构。

6.1.4 生物过程

通过过重分子活动完成的生物学过程，如信号传导，葡萄糖跨膜转运。

6.1.5 基因本体论术语的构成

关于 GO 术语描述见表 2。

**表 9 基因本体论术语**

| 要素名称 | 英文名称 | 描述 | 示例 |
|---|---|---|---|
| 标识符 | Accession | GO 数据库的唯一识别号 | 如：GO:0001591 |
| 名称 | Name | 该 GO 号具体的名字。 | 如：dopamine neurotransmitter receptor activity, coupled via Gi/Go。 |
| 类别 | Ontology | 该术语属于细胞成分，生物过程或分子功能的哪一个。 | 如：Molecular function。 |
| 同义词 | Synonyms | 含义与术语名称紧密相关的替代字词或短语，表示名称与同义词范围所赋予的同义词之间的关系。 | 如：dopamine D2 receptor activity, dopamine D3 receptor activity, dopamine D4 receptor activity。 |
| 次级标识符 | Alternate IDs | 当两个或多个术语的含义相同并且合并为一个术语时，就会出现辅助 ID。所有术语 ID 都会保留下来，因此不会丢失任何信息。 | 如：GO:0001670, GO:0001593, GO:0001592。 |
| 定义 | Definition | GO 的具体文字描述以及信息来用的引用。 | 如：Combining with the neurotransmitter dopamine and activating adenylate cyclase via coupling to Gi/Go to initiate a change in cell activity。 |
| 关系 | Related | 该 GO 与本体中其他 GO 的关系。 | 如：to all genes and gene products annotated to dopamine neurotransmitter receptor activity, coupled via Gi/Go。 |
| 其他 | Other | 其他可选要素。 | 如：Comment, Subset。 |

6.1.6 基因本体论中的关系

GO 以图的形式构建，GO 作为同种的节点，GO 间的关系作为连接。GO 的主要关系描述见表 3。

表 10 基因本体论术语

| 缩写 | 关系 | 表示 | 描述 |
|---|---|---|---|
| i | is a | A i B | 节点 A 是节点 B 的子类型/亚型。 |
| P | part of | A P B | 关系的一部分用于表示整个部分的关系，有当 B 一定是 A 的一部分时，才会在 A 和 B 之间部分关系。 |
| hP | has part | A hP B | 对关系部分的逻辑补充，它从父级的角度代表了“部分-整体”关系。 |
| R | regulates | A R B | 一种过程直接影响另一种过程或质量的表现，即前者调节后者。 |
| R+ | positively regulates | A R+ B | 前者对后者是正向调控。 |
| R- | negatively regulates | A R- B | 前者对后者是负向调控。 |

GO 中三个方面（细胞成分，生物学过程和分子功能）可以用下图 1 的有向无环图表示。

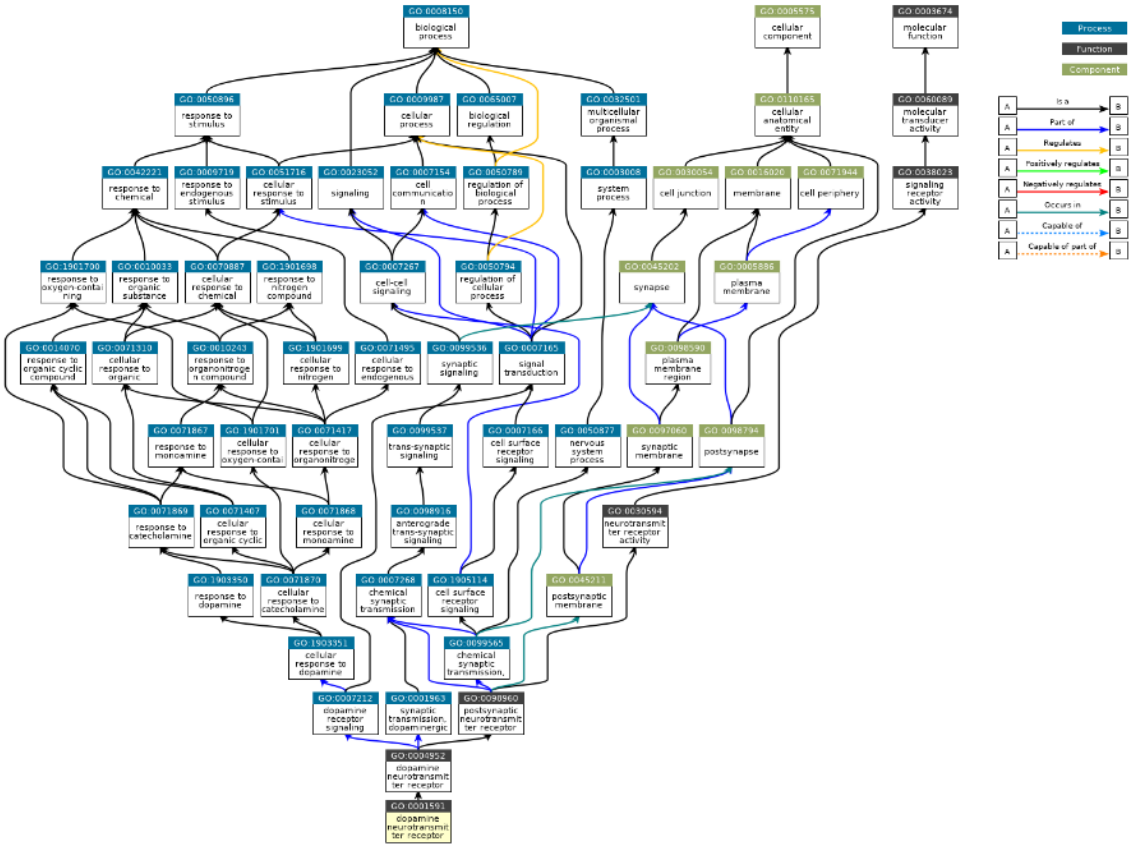


图 1 GO有向无环图