# HPRD数据库处理数据库字段说明

## 数据存储时每个数据表各个字段的名称（含义）

数据表1名称：PROTEIN\_ARCHITECTURE（蛋白质结构）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | HPRD数据库中的标识 | 不一定连续，且可有重复 |
| isoform\_id | String | 标识 | 同一个基因，因启动子，转录起始位点等不同或选择性剪接产生不同的mRNA，进而翻译成有差异的蛋白或多肽，将这些有差异的蛋白或多肽称isoform. |
| refseq\_id | String | 标识 | 美国国家生物信息技术中心（NCBI）提供了具有生物意义上的非冗余的基因和蛋白质序列的RefSeq参考序列数据库。 |
| geneSymbol | String | 基因记号 |  |
| architecture\_name | String | 结构名称 |  |
| architecture\_type | String | 结构名称 |  |
| start\_site | int | 开始位点 |  |
| end\_site | int | 结束位点 |  |
| reference\_type | String | 参比类型 | 具有最佳 氨基酸组成 模式的一种或一组 蛋白 |
| reference\_id | String | 参比标志 |  |

数据表2名称：GENETIC\_DISEASES（遗传疾病）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | HPRD数据库中的标识 |  |
| geneSymbol | String | 基因记号 | 可无 |
| refseq\_id | String | 标识 |  |
| disease\_name | String | 疾病名称 |  |
| reference\_id | int | 参比标志 |  |

数据表3名称：TISSUE\_EXPRESSIONS（组织表达）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | HPRD数据库中的标识 |  |
| refseq\_id | String | 标识 |  |
| geneSymbol | String | 基因记号 | 可无 |
| expression\_term | String | 表达项 |  |
| status | String | 固定值“General”（一般） |  |
| reference\_id | int | 参比标志 |  |

数据表4名称：GENE\_ONTOLOGY（基因本体论）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | 标识 |  |
| isoform\_id | String | 标识 |  |
| refseq\_id | String | 标识 |  |
| geneSymbol | String | 基因记号 | 可无 |
| isoform\_specifity\_status | String | 取值‘General’or’Isoform\_specific’ |  |
| molecular\_function\_term | String | 分子功能项 |  |
| molecular\_function\_reference\_id | String | 分子功能参考标识 |  |
| biological\_process\_term | String | 生物进化 |  |
| biological\_process\_reference\_id | String | 生物进化参考标识 |  |
| cellular\_component\_term | String | 细胞成分 |  |
| cellular\_component\_reference\_id | String | 细胞成分参考标识 |  |

数据表5名称：SEQUENCE\_INFORMATION（序列信息）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | 标识 |  |
| isoform\_id | String | 标识 |  |
| geneSymbol | String | 基因记号 | 可无 |
| nucleotide\_accession | String | 核苷酸加入 |  |
| orf\_start | int | 开始位点 |  |
| orf\_end | int | 结束位点 |  |
| protein\_accession | string | 蛋白质加入 |  |
| protein\_length | int | 蛋白质长度 |  |
| protein\_molecular\_weight | int | 蛋白质分子量 |  |

数据表6名称：HPRD\_ID\_MAPPINGS（标识映射）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | 标识 |  |
| geneSymbol | String | 基因记号 | 可无 |
| nucleotide\_accession | String | 核苷酸加入 | 可无 |
| protein\_accession | int | 蛋白质加入 | 可无 |
| entrezgene\_id | int | 在entrezgene中的标识 |  |
| omim\_id | String | 在omim中的标识 |  |
| swissprot\_id | String | 在swissprot中的标识 |  |
| main\_name | String | 主要名称 |  |

数据表7名称：NUCLEOTIDE\_SEQUENCES（核苷酸序列）

|  |  |
| --- | --- |
| 奇数行 | 标识，名称 |
| 偶数行 | 序列 |

数据表8名称：PROTEIN\_SEQUENCES（蛋白质序列）

|  |  |
| --- | --- |
| 奇数行 | 标识，名称 |
| 偶数行 | 序列 |

数据表9名称：BINARY\_PROTEIN\_PROTEIN\_INTERACTIONS（二进制形式的蛋白质与蛋白质的相互作用）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| interactor\_1\_geneSymbol | String | 相互作用的蛋白质1的基因标记 | 可无 |
| interactor\_1\_hprd\_id | int | 蛋白质1的标识 |  |
| interactor\_1\_refseq\_id | String | 蛋白质1的标识 |  |
| interactor\_2\_geneSymbol | String | 相互作用的蛋白质2的基因标记 | 可无 |
| interactor\_2\_hprd\_id | int | 蛋白质2的标识 |  |
| interactor\_2\_refseq\_id | String | 蛋白质2的标识 |  |
| experiment\_type | String | 体内或是体外进行的实验 |  |
| reference\_id | String | 标识 |  |

数据表10名称：BINARY\_PROTEIN\_NONPROTEIN\_INTERACTIONS（二进制形式的蛋白质与非蛋白质之间的相互作用）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| interactor\_geneSymbol | String | 相互作用的基因标记 |  |
| interactor\_hprd\_id | int | 标识 |  |
| interactor\_refseq\_id | String | 标识 |  |
| non\_protein\_interactor\_name | String | 非蛋白质类物质的名称 |  |
| experiment\_type | String | 体外还是体内 |  |
| reference\_id | String | 标识 |  |

数据表11名称：PROTEIN\_COMPLEXES（蛋白质复合物）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_interaction\_id | String | 标识 |  |
| interactor\_hprd\_id | int | 复合物的hprd标识 |  |
| interactor\_geneSymbol | String | 复合物的基因标记 |  |
| interactor\_refseq\_id | String | 标识 |  |
| experiment\_type | String | 体外还是体内 |  |
| reference\_id | String | 标识 |  |

数据表12名称：PROTEIN\_NOMENCLATURE（蛋白质命名）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| hprd\_id | int | 标识 |  |
| geneSymbol | String | 基因标记 | 可无 |
| refseq\_id | String | 标记 |  |
| all\_nomenclature(separated by semi-colon) | String | 所有的命名（用分号分隔） |  |

数据表13名称：POST\_TRANSLATIONAL\_MODIFICATIONS（翻译后修改）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 含义 | 备注 |
| substrate\_hprd\_id | int | 基质标识 |  |
| substrate\_gene\_symbol | String | 基质基因标记 | 可无 |
| substrate\_isoform\_id | String | 标识 |  |
| substrate\_refseq\_id | String | 标识 |  |
| site | int | 发生位点 |  |
| residue | String | 残留 |  |
| enzyme\_name | String | 作用酶名称 |  |
| enzyme\_hprd\_id | String | 标识 |  |
| modification\_type | String | 修改类型 |  |
| experiment\_type | String | 实验类型 |  |
| reference\_id | String | 标识 |  |