# HPRD数据库分析说明书

## 数据库访问网址（以备检查使用，直接填写即可）

http://www.hprd.org/

## 数据库说明（数据库的介绍，使用范围，包含的数据等内容介绍）

数据库的介绍：The Human Protein Reference Database represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data. HPRD has been created using an object oriented database in Zope, an open source web application server, that provides versatility in query functions and allows data to be displayed dynamically.

使用范围：查询与人类相关的蛋白质数据，如人类蛋白质结构域、翻译后修饰、互作网络和疾病等；查询信号通路；还提供查询蛋白质的磷酸酶基序。

包含的数据等内容介绍：蛋白质名称、相关编号、相关结构、分子类型、结构域、基序、蛋白质修饰类型以及蛋白质的定位。Release9包含制表符分隔的蛋白质-蛋白质互作数据； FLAT\_FILES包含HPRD中蛋白质的全部信息，如：转录后修饰、组织表达、亚细胞定位等，且是制表符分隔的数据；XML也包含蛋白质的全部信息；PSIMI只包含HPRD中全部的蛋白质-蛋白质互作信息，是PSI-MI格式；SINGLE\_PSIMI包含的是单个XML可多行索引的全部的蛋白质-蛋白质互作信息。

## 数据库具体字段说明（说明数据库每个字段包含的意义，如果是XML文件，那么说明每个XML节点中字段的全部含义，即说明数据库中每个字段的对应含义）

数据库具体字段

1. HPRD\_FLAT\_FILES\_041310文件夹——>FLAT\_FILES\_072010文件夹

Number of files = 13

1. PROTEIN\_ARCHITECTURE.txt

hprd\_id,isoform\_id,refseq\_id,

geneSymbol(基因记号),

architecture\_name(结构名称),

architecture\_type(结构类型),

start\_site(开始位点),end\_site(结束位点),reference\_type(参考类型),reference\_id

2. GENETIC\_DISEASES.txt

hprd\_id,geneSymbol,refseq\_id,disease\_name(疾病名称),reference\_id

3. TISSUE\_EXPRESSIONS.txt

hprd\_id,refseq\_id,geneSymbol,expression\_term,status,reference\_id

4. GENE\_ONTOLOGY(本体论).txt

hprd\_id,isoform\_id,refseq\_id,geneSymbol,isoform\_specifity\_status,molecular\_function\_term,molecular\_function\_reference\_id,biological\_process\_term,biological\_process\_reference\_id,cellular\_component\_term,cellular\_component\_reference\_id

5. SEQUENCE\_INFORMATION(序列信息).txt

hprd\_id,isoform\_id,geneSymbol,

nucleotide\_accession(核苷酸),

orf\_start,orf\_end,protein\_accession,protein\_length,protein\_molecular\_weight

6. HPRD\_ID\_MAPPINGS(映射).txt

hprd\_id,geneSymbol,nucleotide\_accession,protein\_accession,entrezgene\_id,omim\_id,swissprot\_id,main\_name

7. NUCLEOTIDE\_SEQUENCES.txt

FASTA Format

8. PROTEIN\_SEQUENCES.txt

FASTA Format

9. BINARY\_PROTEIN\_PROTEIN\_INTERACTIONS.txt

interactor\_1\_geneSymbol,interactor\_1\_hprd\_id,interactor\_1\_refseq\_id,interactor\_2\_geneSymbol,interactor\_2\_hprd\_id,interactor\_2\_refseq\_id,experiment\_type,reference\_id

10. BINARY\_PROTEIN\_NONPROTEIN\_INTERACTIONS.txt

interactor\_geneSymbol,interactor\_hprd\_id,interactor\_refseq\_id,non\_protein\_interactor\_name,experiment\_type,reference\_id

11. PROTEIN\_COMPLEXES.txt

hprd\_interaction\_id,interactor\_hprd\_id,interactor\_geneSymbol,interactor\_refseq\_id,experiment\_type,reference\_id

12. PROTEIN\_NOMENCLATURE.txt

hprd\_id,geneSymbol,refseq\_id,all\_nomenclature(separated by semi-colon)

13. POST\_TRANSLATIONAL\_MODIFICATIONS.txt

substrate\_hprd\_id,substrate\_gene\_symbol,substrate\_isoform\_id,substrate\_refseq\_id,site,residue,enzyme\_name,enzyme\_hprd\_id,modification\_type,experiment\_type,reference\_id

1. HPRD\_Release9\_041310文件夹——>HPRD\_Release9\_062910文件夹

Number of files = 1

1. BINARY\_PROTEIN\_PROTEIN\_INTERACTIONS.txt

This file stores all the binary protein protein interactions.

This file contains the following columns:

1. Interactor 1 Gene symbol

2. Interactor 1 HPRD id

3. Interactor 1 RefSeq id

4. Interactor 2 Gene symbol

5. Interactor 2 HPRD id

6. Interactor 2 RefSeq id

7. Experiment type (in vivo, in vitro and yeast 2-hybrid)

8. Pubmed id

1. xml文件格式

<HPRD3r> HPRD（human protein reference database）数据库是包含蛋白质注释、PPI、转录后修饰和亚细胞定位等多种信息的综合数据库

<protein > 蛋白质，包含三个属性（isoform,version,id）

<titles>标题集

<title>标题

<alt\_title>替代标题

<omim>OMIM “0nline Mendelian Inheritance in Man”的简称，即在线《人类孟德尔遗传》，是一个关于人类基因和遗传紊乱的数据库（持续更新）。主要着眼于遗传性的基因疾病，包括文本信息和相关参考信息、序列纪录、图谱和相关其他数据库。

<gene\_symbol>基因符号

<gene\_map\_locus>基因图位点

<title>标题

<isoform>isoform：同种型，同等型，同工型。由来自一个基因的mRNA前体因选择性剪接而产生多种mRNA， 并翻译出的不同蛋白质，称为同源异构体(isoform)。

<seq\_entry >条目序列 , 包含一个属性（source）

<entry\_cdna>cDNA条目

<entry\_protein>蛋白质条目

<molecular\_weight>分子量

<entry\_sequence>entry序列,包含一个属性（length）

<cdna\_sequence >cDNA序列，即经过反转录得到的与RNA序列互补的DNA序列,包含连个属性（start,end）

<cdna\_utr5/>UTR（Untranslated Regions)即非翻译区，是信使RNA（mRNA）分子两端的非编码片段。转录而成的mRNA序列除去编码生成多肽的coding squence，在5‘端和3’端还有一些不翻译的序列 5'-UTR从mRNA起点的甲基化鸟嘌呤核苷酸帽延伸至AUG起始密码子,包含连个属性（start,end）

<cdna\_coding />cDNA编码,包含两个属性（start,end）

<cdna\_utr3 />UTR（Untranslated Regions)即非翻译区，是信使RNA（mRNA）分子两端的非编码片段。转录而成的mRNA序列除去编码生成多肽的coding squence，在5‘端和3’端还有一些不翻译的序列 3'-UTR从编码区末端的终止密码子延伸至多聚A尾巴（Poly-A）的末端,包含连个属性（start,end）

<protein\_sequence >蛋白质序列,包含一个属性（length）

<protein\_domain\_architecture>蛋白质结构域

<domain >域,包含四个属性（domain\_source，start,end,type）

<title>题目

<modifications>修饰集合

<modification >修饰,包含一个属性（type）

<ptm\_site>翻译后修饰(PTM)位点

<ptm\_residue>翻译后修饰残基

<exp\_type>实验类型

<expressions>实验集合

<expression>实验

<title>标题

<pubmed>pubmed文献数据库索引

<functions>功能

<molecule\_class>分子类

<title>题目

<molecular\_function>分子功能

<title>标题

<go\_id>基因本体（Gene Ontology，GO）是一个在生物信息学领域中广泛使用的本体。它主要包括三个分支: 生物过程、分子功能和细胞组件。

<biological\_process>生物过程

<title>标题

<go\_id>基因本体（Gene Ontology，GO）是一个在生物信息学领域中广泛使用的本体。它主要包括三个分支: 生物过程、分子功能和细胞组件。

<entrySet> entry集合

<entry>是UniProt给每个蛋白质赋予的独一无二的ID号

<source>资源

<names>名称

<shortLabel>短标签

<fullName>全称

<bibref>参考文献

<xref>外部参考,包含两个属性（db,id）

<primaryRef />主要参考,包含两个属性（db,id）

<secondaryRef/> 次要参考

<availabilityList>可用性列表

<availability >可用性,包含一个属性（id）

<experimentList>实验列表

<experimentDescription >实验说明,包含一个属性（id）

<bibref>参考文献

<xref>外部参考

<primaryRef />主要参考,包含五个属性（db,dbAc,id,refType,refTypeAc）

<secondaryRef />次要参考,包含五个属性（db,dbAc,id,refType,refTypeAc）

<interactionDetectionMethod>交互检测方法

<names>名称

<shortLabel>短标签

<fullName>全称

<xref>外部参考

<primaryRef />主要参考,包含五个属性（db,dbAc,id,refType,refTypeAc）

<interactorList>交互者列表

<interactor >交互件 | 相互作用因子，包含一个属性（id）

<names>名称

<shortLabel>短标签

<alias >别名,包含两个属性（type,typeAc）

<xref>外部参考

<primaryRef />主要参考,包含五个属性（db,dbAc,id,refType,refTypeAc）

<secondaryRef/> 次要参考,包含三个属性（db,dbAc,id）

<interactorType>交互者类型

<names>名称

<shortLabel>短标签

<fullName>全称

<xref>参考文献

<primaryRef />主要参考 ,包含三个属性（db,dbAc,id）

<organism >蛋白质的种属来源,包含一个属性（ncbiTaxId）

<names>名称

<shortLabel>短标签

<fullName>全称

<sequence>序列

<interactionList>相互作用列表

<interaction >相互作用，包含一个属性（id）

<experimentList>实验列表

<experimentRef>实验参考

<participantList>参与者列表

<participant >参与者,包含一个属性（id）

<interactorRef>交互者参考

<external\_links>外部链接

<SwissProt>SwissProt数据库：Swiss-Prot (reviewed) 是一个高质量人工注释且非冗余的蛋白序列数据库。其中包含各类实验结果、计算得到的特征信息和文献信息等。

<EntrezGene>EntrezGene数据库：美国生物信息技术中心（NCBI)的字库Entrez Gene主要提供人类及其他动物等基因信息。

<PDB>PDB(Protein Data Bank)数据库 ：生物大分子结构数据库，提供蛋白质、核酸等生物大分子的三维结构数据、序列详细信息、生化性质等。

<author>作者

<annotator>注释者

<last\_updated>最近更新

## 数据库进行更新时，数据更新方式（全库更新或增量更新，更新时数据是否包含1、新增数据；2、修改数据、3、删除数据）

## 数据下载方式，数据更新后的下载地址（下载地址每次有变化或者下载的规律。下载下来的文件是单个文件还是压缩包。数据库更新的频率或每次更新的日期。）

数据库下载地址：<http://www.hprd.org/download>（需要填写一些信息）

下载下来的是压缩包

更新：We hope to keep it updated with the assistance of the entire biomedical community. Any licensing fee, if generated, will be used to annotate HPRD better and to add more entries and features.In the future, we will depend on the biology community for such updates in addition to our own efforts.