

蛋白质数据库

先说重点内容：

小结一下数据库的选择：UniProt Proteomes 是第一优先级，除了人和小鼠可以用 Swiss-Prot 更好以外，其余所有物种都优先用 Proteomes，第二优先级 GenBank/NCBI Inr，用于少数 UniProt 没有收集到的物种。

Tips:

Swiss-Prot 是经过人工注释和 review 的数据库，它只有 55 万种蛋白。但是除了人类和小鼠蛋白质数据库因为研究得很广泛比较全面以外，剩下的各个物种都不是很全面。因此建议除了人类和小鼠外，别的物种不要直接采用 swiss-prot 搜库。

以下为正文

我们简单介绍一些蛋白质数据库的知识。先来个总表，大家感受一下

蛋白数据库及其注释信息的构建



1：序列数据库

目前来说，用的最多的是UniProt KB，该数据库来自欧洲生物信息学中心。其次是美国的NCBI Genebank。这两个数据库搜集了全世界已公布的所有物种的蛋白质序列。如果实在搜不到结果，还可以用EST标签或者自己去测序，只是自己测序无法保证蛋白的完整程度。

2：注释数据库

鉴定到蛋白只是万里长征的第一步，后面我们还需要对蛋白进行注释，比如我们最常用到的Gene Ontology。人类蛋白数据库已经注释得很完整，而有的物种注释不够完整或者说注释得比较差的情况下，则需要通过同源性序列来间接注释。

3：蛋白相互作用数据库

当需要进一步研究蛋白的功能及作用机理时，常常需要了解蛋白-蛋白或蛋白-小分子相互作用，有很多收集蛋白相互作用的数据库可以供我们搜索，或者绘制互作网络。

4：生物通路分析数据库

比如大名鼎鼎的KEGG等工具，还有一些有偏好性的数据库，比如专门针对代谢通路的BioCyc，或者针对人类（及大鼠和小鼠）物种的IPA等。

5：蛋白质组学数据库

当我们完成了从搜库、注释，到机理分析的一系列功能，并完成了生物学实验验证，打算发表文章了。有一些蛋白质组学领域的杂志，比如JPR、MCP等，会要求我们将数据结构上传到指定的数据库中，用于共享或是同行的质量审查。目前来说用的最多的是ProteomeXchange，ProteomeDB，和iProx这三个数据库。iprox是中国国家蛋白质中心建立的。另外一些蛋白质组学相关的数据库，以及发表在CNS上的大规模数据，有一些组织也会将其搜集起来，做人的human protein atlas，比如GeneCards就是整合得很好的综合性数据库，我们可以在其中查到别人做过的详细结果数据。

这么多种类繁杂的数据库，相互之间的数据信息有怎样的联系呢？下面这张图告诉你

答案：

各种各样的信息，各种和生化以及组学相关的数据库都可以整合到 NCBI 中。NCBI 支持的数据格式包括 NCBI GI、GenBank ID、RefSeq ID，以及 Entrez ID 等。

NCBI 的 NCBIInr 非冗余数据库是搜库时经常会用到的，但它的问题在于 NCBI 内部数据的一致性比较差，它搜集了各种来源的数据，格式都不一样，后续会发现，搜集到 NCBIInr 之后，同一个基因编码的蛋白会搜到好多个版本。

2016 年，NCBI 将 gi 号取消了，换成了 GenBank ID，此过程十分艰难，很多软件都要对其进行相应的转换，也给使用者带来了很大不便。因此个人建议，还是先在 UniProt 库里搜索。如果从 UniProt 里实在找不到的序列信息，再去 NCBIInr 里搜索。

Tips: 虽然 UniProt 主要搜集的是蛋白信息，但是它与相当多的注释数据库，如 GO，KEGG 等等，都有交叉合作。因此 UniProt 中的注释信息是相当完整的。

但是 NCBI 的优势也是非常明显的，就是它的数据信息非常全面！从下图可以看出，在过去的 7 年时间里，NCBI 包含的核酸序列、蛋白序列和基因信息均有爆发式的增长。这归功于近年来基因组和转录组技术的发展。其中很多是中国人做出的贡献。

NCBI

■ 数据库统计

分类	2009.5	2016.10	增长	增长比例
Pubmed	18832968	26608697	7775729	41%
Nucleotide	76750026	220008176	143258150	187%
Protein	26369028	329407218	303038190	1149%
Gene	5798947	25687392	19888445	343%
UniGene	3633531	6473284	2839753	78%

刚刚我们也提到了 NCBI 的问题，那么它的缺点对我们搜库有什么影响呢？

举个例子。

比如有一次我们做某种橘子的蛋白鉴定，在 NCBI 中搜索，如果用 NCBIInr(非冗余)来搜，你会得到 88138 条蛋白序列，但其中有 21% 的序列是完全一样的，原因就是其包含的数据来源太多了！如果我们用 GenBank，就会发现只有 15% 的冗余（GenBank 也不是单一来源的数据库，它自己也有好几个注释序列的来源）。

而当我们选用 UniProt 的话，发现结果里没有冗余！这就是 UniProt 的好处，帮助我们进行了前期蛋白数据库的过滤和准备。这就是我们推荐优先使用 UniProt 的原因。

事实上，现在用 UniProt 的人越来越多了。它是目前世界上最大最完整的蛋白数据库，其来源非常多，比如有 GenBank，EMBL-Bank，DDBJ 等的 coding sequences 都会成为其收集来源。

既然 UniProt 这么好用，我们再来介绍一下它是如何使用。

UniProtKB

■ UniProtKB是目前最大、最完整的蛋白序列数据库 <http://www.Uniprot.org>



首先，我们得确认一下所要搜索的物种的拉丁文名称，比如说猪，因为是很常见的物种，所以在其拉丁文名后，包含了其英文名 pig。但你如果用 sus scrofa 来搜索会得到最准确的结果，否则会得到大量的候选。

Sus scrofa (Pig)

Overview

Proteome name: Sus scrofa - Reference proteome

Proteins: 26,101

Proteome ID: UP00000227

Strain: Duroc

Taxonomy: 9623 - Sus scrofa

Last modified: July 9, 2016

Genome assembly: GCA_000003025.4

Sus scrofa, the domestic pig, is a member of the Artiodactyla (even-toed ungulates). First domesticated around 10,000 years ago, they are an important agricultural species worldwide. In addition, pigs are useful as a non-primate biomedical model organism, especially for the study of complex traits (e.g. obesity, cardiovascular disease, and arthritis) and for xenotransplantation experiments.

The first draft genome was released in 2012 by the Swiss Genome Sequencing Consortium, using sequence derived from a single Duroc sow (Duroc 2-14). Assembly 10.2 contains approximately 2.6 Gb of sequence and 21,300 protein-coding genes.

Components

Download | View all proteins

Component name	Genome Accession(s)	Proteins
Chromosome 1	CH000812	2084
Chromosome 2	CH000813	2154
Chromosome 3	CH000814	1417

搜索完毕后，我们可以得到具体数据，比如蛋白数量 26000 种，编码基因 21000 种，这个时候我们可以有个预判：猪的蛋白质组相对来说是比较完整的。在该搜索页面中，有 download 选项，点击后可以进入相应的下载界面，下载到本地，导入搜库软件中，

就可以使用了。

Tips

目前在UniProt Proteomes数据库里，有reference proteome（数据库里会用R标签来表示）的物种目前有5862种。另外51999种物种有proteome但是没有reference，说明其数据相对来说还不够全面。

注：此贴是转载贴，非原创。仅供内部学习使用。