Lab1 探索NCBI

由美国国家医学图书馆和国立卫生研究院维护的国家生物技术信息中心（NCBI）是全球最重要的生物数据资源和存储库之一。 这个在线资源提供了广泛的来自各行各业的越来越多的遗传信息资源。在这个网站中，可以对从病毒到人类的整个基因组进行编译，组织和交叉引用，因此浏览基因组几乎就像在网上冲浪一样容易。

我们将在本章学习如何使用NCBI。

1. NCBI的主要门户为*search NCBI*，我们需要先打开NCBI的主页，网址为www.ncbi.nlm.nih.gov。改网页提供了NCBI的所有数据库和资源，可以通过查阅NCBI手册http://www.ncbi.nlm.nih.gov/books/NBK21101/来获取更多信息。

2. 接下来前往*Search NCBI*界面（一般被称为*GQuery/Entrez*），在页面搜索界面中输入search即可，可以输入***bacteria***进行尝试，我们将会看到好几百万个结果，因此我们需要缩小结果的范围。

3. 因此在使用该数据库时，我们可以输入DNA片段或我们感兴趣的蛋白质（编码）来缩小搜索范围。在本例中我们使用拟南芥，一种植物学的模式植物作为例子，该基因是一种小型开花植物，就像植物世界的果蝇一样，因为它的生命周期相对较快，并且需要的生长空间很小。 我们要查阅的基因的蛋白质产物以登录号NP_001318308记录，它是一种E3连接酶，参与蛋白质的泛素化，这是其降解的信号。

4. 返回到Search NCBI门户页面，然后尝试更集中的搜索。 使用我们将在下面显示的在GenBank Field Qualifiers（GeneBank字段限定符）中使用的基因序列相关的搜索词（完整的限定词列表在附录1中提供）。 尝试下面显示的四种不同搜索，然后查看找到的数字记录，特别是“蛋白质”记录。最后，必须使用大写布尔运算符例如AND / OR / NOT。 最终，你可以使用的最具体的搜索项目是登录号：

   gene keywords：
   e.g. **ubiquitin-protein ligase**
   gene keyword AND organism:
   e.g. **ubiquitin-protein ligase AND Arabidopsis thaliana**
   gene keyword [PROT] AND organism [ORGN]:
   e.g. **ubiquitin-protein ligase [PROT] AND Arabidopsis thaliana [ORGN]**
   Accession or GI number
   e.g. NP_001318308

5. 通过“搜索NCBI”门户页面搜索我们感兴趣的登录号（例如，上面的NP_001318308）。它应该给你唯一的一个蛋白序列命中。 单击它（它是超链接），以便获得其完整的GenBank描述（也可以单击页面顶部的“ armadillo /beta-catenin repeat protein [Arabidopsis

thaliana]”链接，因为NCBI系统会识别出该链接 您已经输入了蛋白质标识符，因此在结果的数值概览上方提供了一些摘要信息）。

6. 请注意文本中的所有超链接。 它看起来很杂乱，但实际上很简单。 例如，有关分类信息，请单击“**SOURCE ORGANISM**”超链接。 一些记录具有指向主要出版物的链接，该出版物最初是在**PUBMED**数字超链接中引用该序列的（上述示例中不是这种情况，但是该序列有PubMed参考）。单击不同的链接，然后查看您发现的内容。

Box1. Accession Numbers, Version Numbers, and GI Numbers
（登记号，版本号与GI号）
登记号是特定序列记录的唯一标识符。 将登记号分配给特定记录，并永久保留在该记录中。 换句话说，登录号跟踪特定的记录，即使记录中的信息根据作者的请求而更改（例如，如果提供了更好的注释或更完整的序列），则登记号也不会更改。 入藏号通常是字母和数字的组合，例如单个字母后跟五个数字（例如U12345）或两个字母后跟六个数字（例如AF123456）。
版本号紧随登录号之后，指示该条目的修订历史记录，从1开始，并随着每个修订版本的增加而增加。 标准格式为*Accession.Version*。
GI号（GenInfo标识符-有时用小写的“ gi”表示）只是一系列数字，直到最近，这些数字都连续分配给NCBI处理的每个序列记录。 标识符的GI系统与*Accession.Version*系统并行运行。 因此，如果DNA或蛋白质序列发生任何变化，它将收到一个新的GI号。
示例：将新条目提交给GenBank时，会为其分配一个登录号（例如AF000001）。 由于这是第一个版本，因此该附件将附加有“ .1”，因此看起来像AF000001.1。 同时提供了GI号（例如GI：1234567）。 现在想象一下，最初提交记录的研究人员想要更新信息。 更新的记录将保留相同的登录号，但版本号将增加（AF000001.2）。 新记录将被赋予一个全新的GI号（例如GI：9876543）。
为什么这很重要？ 登记号将始终为您提供记录的最新信息，而GI/*Accession.Version*将始终将您带到特定记录。 有时候，您需要最新的信息，而有时候，您想要从特定的时间点指向特定的信息（例如，您进行分析的特定记录），即使有更多的信息被使用。 随后添加。 请注意，自2016年9月起，NCBI开始逐步淘汰GI号的使用。 现在建议使用Accession.Version表单而不是GI号来访问特定记录。
Ps：在NCBI主页的左下方，找到“ NCBI帮助手册”链接。点击它。 然后访问“ Entrez Help”部分即可查看帮助索引。

7. 返回GenBank记录，然后单击实际序列上方的**CDS**链接。

8. 返回GenBank记录并检查右下角的“**Related Information**”部分。这使您可以直接链接到具有此查询信息的其他数据库。 找到**Gene**链接。

9. 从“**Related Information**”菜单中选择“**Gene**”。 这是NCBI很好的入门资源。 滚动浏览不同的部分并回答下列问题：

a. 您的基因在基因组中的位置在哪里？提示：将鼠标悬停在“Genomic regions, transcripts, and products”部分的绿色条上；绿色条代表序列查看器中的基因
b. 您在这个基因中看到多少个外显子？ 提示：有多少个绿色盒子？
c. 围绕它的基因的名称是什么（即它的“Genomic context”是什么）？
d. 它有任何保守域(Conserved Domains)吗？ 它们叫什么？ （提示：使用“Related Information”链接到“**Gene**”页面右侧的保守域）
e. 探索了保守域后，返回“基因”页面。 该基因涉及什么生物学过程（基因本体论术语）（向下滚动页面查看）？

10. 在Gene界面上，还有**Additional links**可以进一步检查基因的结构，功能和系统发育关系。 右侧的导航侧栏上有一个“Additional links”超链接，它将带您到页面底部，在大多数基因中都找到它们。 单击[+]Gene LinkOut以查看它们。我们可以探索PUB12数据互连和显示的各种方式。使用“**Related Information**”链接，您可以找到与此基因相关的任何出版物吗？ 基因表达数据呢？ 下一页显示了相应mRNA的相关“ RefSeq RNA”记录（NCBI的RefSeq旨在为许多模型生物提供规范的“reference”序列-基因组，mRNA，CDS，蛋白质等）。

Box2. Helpful Hints for NCBI searches
关于NCBI search的搜索提示
在大多数NCBI搜索页面上（奇怪的是，搜索NCBI除外），单击搜索框下方的“保存搜索”或“创建警报”。 注册一个帐户并保存搜索。 您还可以使用“历史记录”标签和其中列出的搜索编号来组合以前的搜索，以及通过注册“我的NCBI”帐户来保存搜索，因此以后不必继续重做相同的搜索。

附录：GenBank Field Qualifiers| GenBank限定词列表位于下面的网页中
http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options