一级核酸数据库：GenBank 原核生物核酸序列（1）

这一节我们来看一级核酸数据库，他主要包括三大核酸数据库和基因组数据库。三大核酸数据库包括 NCBI 的 Genbank，EMBL 的 ENA 和 DDBJ，它们共同构成国际核酸序列数据库。三大核酸数据库，美国一个，欧洲一个，亚洲一个。美国的 Genbank 由美国国家生物技术信息中心 NCBI 开发并负责维护。NCBI 隶属于美国国立卫生研究院 NIH。欧洲核苷酸序列数据集 ENA 由欧洲分子生物学研究室 EMBL 开发并负责维护。亚洲的核酸数据库 DDBJ 由位于日本静冈的日本国立遗传学研究所 NIG 开发并负责维护。Genbank，EMBL 与 DDBJ 共同构成国际核酸序列数据库合作联盟 INSDC。通过 INSDC，三大核酸数据库的信息每日相互交换，更新汇总。这使得他们几乎在任何时候都享有相同的数据。

我们以 NCBI 的 Genbank 为例，教你如何解读一级核酸数据库。我们将分别浏览一个原核生物的基因和一个真核生物的基因。为此，请先跟我复习一下原核生物与真核生物基因的不同之处。原核生物基因组小，真核生物基因组大。原核生物基因密度高，1000 个碱基里就有 1 个基因，而真核生物基因密度低，比如人，要 10 万个碱基才有 1 个基因。与此对应，原核生物编码区含量高，而真核生物低。此外，原核生物的基因是呈线性分布的，而真核生物的基因是非线性的，因为翻译蛋白质的外显子被内含子分隔开来。也就是真核生物的 mRNA 要经历剪切的过程，剪切后的成熟 mRNA 才能进行翻译。这是原核生物和真核生物基因的最大区别，即，原核生物没有内含子，真核生物有内含子。这个巨大的区别，将导致两种基因在数据库中不同的存储及注释方式。

我们首先来看一条原核生物的 DNA 序列，它是编码大肠杆菌 dUTPase 的基因，在 Genbank 里的数据库编号是 X01714。从 NCBI 的主页（http://www.ncbi.nlm.nih.gov/）选择 Genbank 数据库。Nucleotide 数据库就是 Genbank 数据库，然后在搜索条中直接写入这条序列对应的数据库编号 X01714，点击“搜索”。结果返回编号为 X01714 的序列在 Genbank 中详细记录。从这条记录的标题我们得知，dUTPase 是脱氧尿苷焦磷酸酶，编码他的基因叫 *dut* 基因，所属物种是大肠杆菌。下面是关于这个基因的详细注释，我们逐条浏览一下：

LOCUS 这一行里包括基因座的名字，核酸序列长度，分子的类别，拓扑类型，原核生物的基因拓扑类型都是线性的，最后是更新日期。

DEFINITION 是这条序列的简短定义，也就是前面看到的标题。

ACCESSION 就是在搜索条中输入的那个数据库编号，也叫做检索号，每条记录的检索号在数据库中是唯一且不变的。即使数据提交者改变了数据内容，Accession 也不会变。你会发现，这条记录里，Accession 和 Locus 是一样的。这是因为这个基因在录入数据库之前并没有起名字，因此录入数据的时候便将检索号作为了基因的名字。但是有些基因，在录入数据库之前已经有了自己的名字，那么这些基因所对应的 Accession 和 Locus 就不一样了。你可以这样理解，Locus 是一个同学的真实姓名，而 Accession 是这个同学的学号。同一个人在不同的学校里会有不同的学号，而名字只有一个。基因也是一样，同一个基因在不同的数据库中会有不同的检索号，而基因的名字只有一个。

Version 版本号和 Locus，Accession 长得差不多。版本号的格式是“检索号点上一个数字”。版本号于 1999 年 2 月由三大数据库采纳使用。主要用于识别数据库中一条单一的特定核苷酸序列。在数据库中，如果某条序列发生了改变，即使是单碱基的改变，它的版本号都将增加，而它的 Accession 也就是检索号保持不变。比如，版本号由 U12345.1 变为 U12345.2，而检索号依然是 U12345。版本号后面还有个 **GI** 号。GI 号与前面的版本号系统是平行运行的。当一条序列改变后，它将被赋予一个新的 GI 号，同时它的版本号将增加。