

# 核酸/蛋白序列本地管理工具-使用指南

---



## 核酸/蛋白序列本地管理工具-使用指南

简介

版本与使用库说明

本文件夹文件说明

具体使用方法介绍

1.主页介绍

2.本地序列库的使用方法

(1) 添加新序列

(2) 删除序列

(3) 下载目标序列

(4) 清空数据库

3.NCBI序列爬取工具使用方法

## 简介

---

这是一个本地的管理自己常用的核酸或蛋白的序列存储库，可以手动添加，也可以使用NCBI抓取工具对目标序列进行获取，减少反复浏览NCBI数据库的次数；保存自己常用的序列数据也能够方便后续的分析。

## 版本与使用库说明

---

工具使用的软件版本为：

- Python 3.9.13

使用的相关库为：

- tkinter
- urllib
- re

- sleep

上述库都是Python标准库，只要安装了Python便无需单独下载！

## 本文件夹文件说明

---

该文件所处文件夹的个文件说明如下：

- 说明文件：
  - README.pdf: 即 本说明文件
- 装饰图片：
  - icon.ico: 工具对话框左上角呈现的图标
  - 背景.gif: 工具主页图片
- 必要存储文件：
  - table.txt: 存储序列库的信息，用于存储和展示，包括序列类型、注册号、名称、长度、来源（初始可为空文件）
  - seq.txt: 存储序列的详细信息，用于存储和下载，包括序列类型、注册号、FASTA序列内容（初始可为空文件）
- 主程序：
  - main.py: 工具主程序，双击即可运行

为保证运行正确，上述文件应当在同一文件夹下！

## 具体使用方法介绍

---

### 1.主页介绍

双击main.py，进入主页：



双击后会弹出主页（上图）及黑色的Python运行框。

一共有三个选项，其中“本地序列库查询”和“从NCBI下载序列”两个按钮会在下面两节详细介绍；

“退出”按钮可以结束该程序。

## 2.本地序列库的使用方法

按1中的方法进入主页后，点击“本地序列库查询”，进入序列信息数据库：

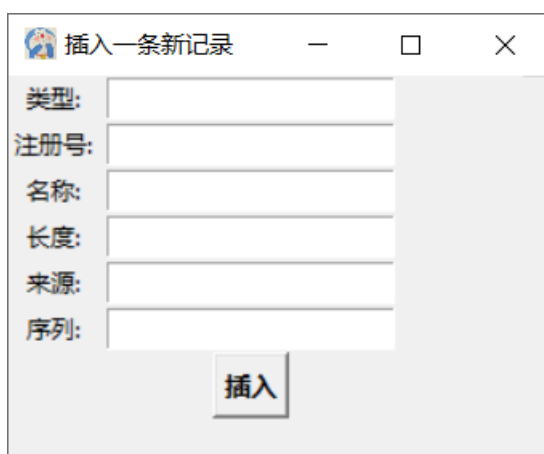


可以看到此时数据库里面没有东西。

在页面顶部工具栏有四种操作方法，依次介绍：

### (1) 添加新序列

我们点击“添加新序列”，会弹出以下窗口：

A screenshot of a dialog box titled "插入一条新记录" (Insert a new record). The dialog box has a light gray background and a white border. It contains six input fields, each with a label to its left: "类型:" (Type), "注册号:" (Registration Number), "名称:" (Name), "长度:" (Length), "来源:" (Source), and "序列:" (Sequence). At the bottom center of the dialog box is a button labeled "插入" (Insert).

这是手动添加一条序列信息，需要我们手动输入，比如：

类型:	Nucleotide
注册号:	AY230262
名称:	Homo sapiens NANOG mRNA, partial cds.
长度:	918
来源:	Homo sapiens (human)
序列:	AACCTGAAGACGTGTGA

插入

上图中填入的内容为：

- 类型: Nucleotide
- 注册号: AY230262
- 名称: Homo sapiens NANOG mRNA, partial cds.
- 长度: 918
- 来源: Homo sapiens (human)
- 序列:
  - ATGAGTGTGGATCCAGCTTGTCCCCAAAGCTTGCCTTGCTTTGAAGCATCCGACTGTAAA  
GAATCTTCAC  
CTATGCCTGTGATTTGTGGCCTGAAGAAACTATCCATCCTTGCAAATGTCTTCTGCTGA  
GATGCCTCA  
CACGGAGACTGTCTCTCCTCTTCCTTCTCCATGGATCTGCTTATTCAGGACAGCCCTGAT  
TCTCCACC  
AGTCCCAAAGGCAAACAACCCACTTCTGCAGAGAAGAGTGTGCAAAAAAGGAAGACAA  
GGTCCCGGTCA  
AGAAACAGAAGACCAGAACTGTGTTCTTCCACCCAGCTGTGTGTACTCAATGATAGAT  
TTCAGAGACA  
GAAATACCTCAGCCTCCAGCAGATGCAAGAACTCTCCAACATCCTGAACCTCAGCTACAA  
ACAGGTGAAG  
ACCTGGTTCCAGAACCAGAGAATGAAATCTAAGAGGTGGCAGAAAAACAACCTGGCCGAA  
GAATAGCAATG  
GTGTGACGCAGAAGGCCTCAGCACCTACCTACCCAGCCTTTACTCTTCCTACCACCAGG  
GATGCCTGGT  
GAACCCGACTGGGAACCTTCCAATGTGGAGCAACCAGACCTGGAACAATTCAACCTGGA  
GCAACCAGACC  
CAGAACATCCAGTCCTGGAGCAACCACTCTGGAACACTCAGACCTGGTGCACCCAATCC  
TGGAACAATC  
AGGCCTGGAACAGTCCCTTCTATAACTGTGGAGAGGAATCTCTGCAGTCCTGCATGCAGT  
TCCAGCCAAA  
TTCTCCTGCCAGTGACTTGGAGGCTGCCTTGAAGCTGCTGGGGAAGGCCTTAATGTAA  
TACAGCAGACC  
ACTAGGTATTTTAGTACTCCACAAACCATGGATTTATTCCTAACTACTCCATGAACATGCA

ACCTGAAG  
ACGTGTGA

注意：

- 类型只能是Nucleotide或Protein，因为该工具目的只存储这两个库的序列信息
- 注册号对应NCBI上的ACCESSION
- 名称对应NCBI上的DEFINITION
- 来源对应NCBI上的SOURCE
- 序列对应NCBI上的FASTA序列信息，不要标题，只输入序列

收到添加成功的信息后，我们可以看到界面更新了我们新加入的这一条序列的信息：

数据库				
添加新序列 删除序列 下载目标序列 清空数据库				
类型	注册号	名称	长度	来源
Nucleotide	AY230262	Homo sapiens NANOG m	918	Homo sapiens (human)

这就说明添加成功了！

当然，你也可以直接打开table.txt和seq.txt查看相关内容！

## （2）删除序列

作为**示例**，我们这样添加信息到库中：

数据库				
添加新序列 删除序列 下载目标序列 清空数据库				
类型	注册号	名称	长度	来源
Nucleotide	AY230262	Homo sapiens NANOG m	918	Homo sapiens (human)
Nucleotide	AA111111	x	x	x
Nucleotide	AA222222	x	x	x
Protein	AA333333	x	x	x
Protein	AA444444	x	x	x
Protein	AA555555	x	x	x

(仅作参考!)

如果要删除序列，可点击“删除序列”，进入删除序列界面：

删除记录

请输入要删除的序列注册号:

(如果有多条,请务必以英文','间隔!!!)

删除

这里可以接受通过注册号删除，而且可以同时删除多条！

**注意，在同时删除多条时，一定要用英文逗号间隔！**

示例：

删除记录

请输入要删除的序列注册号:

AA111111,AA444444

(如果有多条,请务必以英文','间隔!!!)

删除

点击“删除”：

数据库				
添加新序列 删除序列 下载目标序列 清空数据库				
类型	注册号	名称	长度	来源
Nucleotide	AY230262	Homo sapiens NANOG m	918	Homo sapiens (human)
Nucleotide	AA222222	x	x	x
Protein	AA333333	x	x	x
Protein	AA555555	x	x	x

删除成功，没有相关信息了！

### (3) 下载目标序列

如果想要下载自己库中的序列，可以使用“下载目标序列”菜单。点击后进入：

下载序列

请输入要下载的序列注册号:

(如果有多条,请务必以英文','间隔!!!)

下载路径:

选择路径

下载

可以同时下载多条序列，但是需要以英文逗号分隔！

如图示例，我下载序列AA2222和AA5555到C:/Users/cien/Desktop/download.txt文件中：（需要修改为自己的路径！）

下载序列

请输入要下载的序列注册号:

AA222222,AA555555

(如果有多条,请务必以英文','间隔!!!)

下载路径:

C:/Users/cien/Desktop/c

选择路径

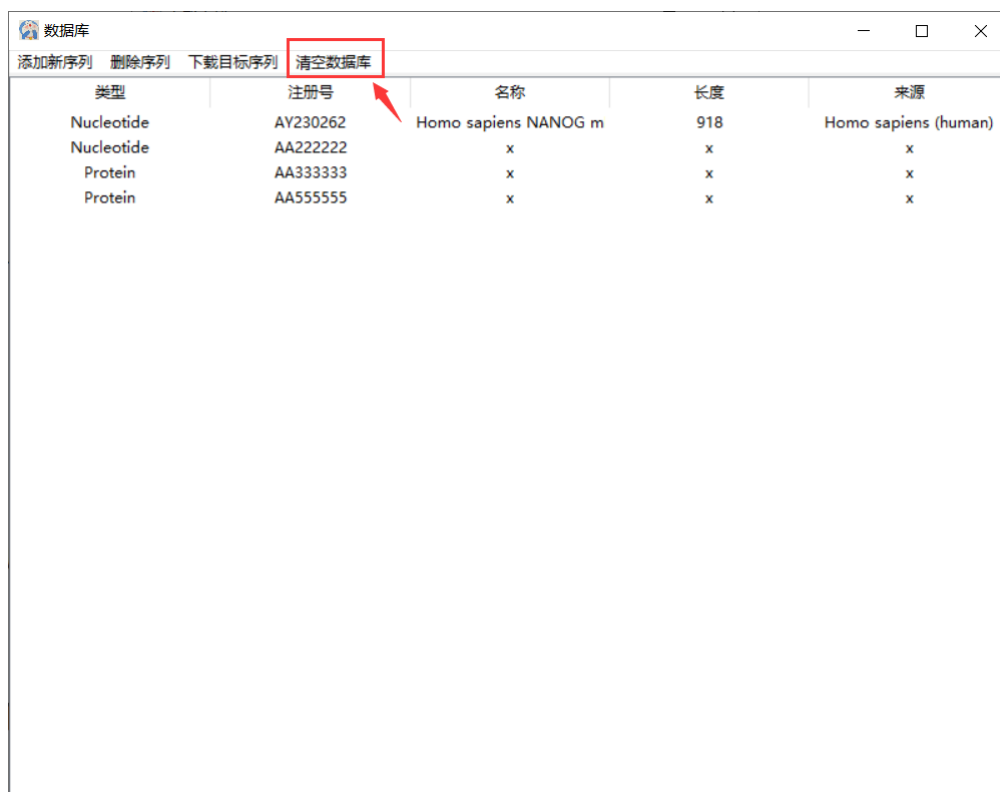
下载

点击“下载”，就可以在对应的文件中看到下载的信息了：

```
download.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
> AA222222 x
SNLKCNCNLCNSJNCLNSLKNC
> AA555555 x
LSJCBIKLCBNLJSBC
```

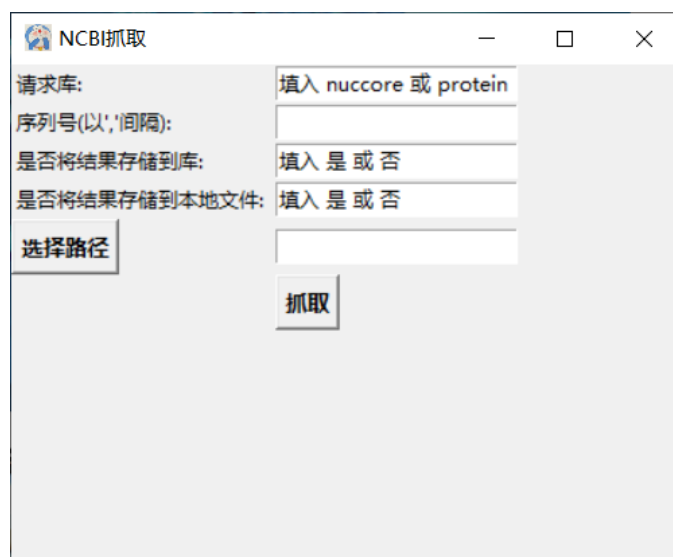
#### (4) 清空数据库

点击“清空数据库”，可以清除页面显示和本地两个文件table.txt和seq.txt：



### 3.NCBI序列爬取工具使用方法

点击主页的第二个按钮“从NCBI下载序列”可以进入NCBI序列下载工具：





可以看到，需要我们提交一些信息，从而帮助我们自动去NCBI抓取我们想要的序列信息！

填写说明：

- 请求库只能是核酸核心库nucore和蛋白库protein，拼写一定要方框中书写所示！
- 序列号可以一个或多个，但是必须以英文逗号间隔！
  - 可以一次请求多个，但是必须隶属于同一个库！
- “是否将结果存储到库”：是否把爬取到的内容放在刚刚我们设计的本地库中
- “是否将结果存储到本地文件”：是否要把爬取的FASTA文件整合为一个文件存储到本地计算机
  - 如果选择了这一项，一定要给出文件路径！！！

下面是一个填写示例：

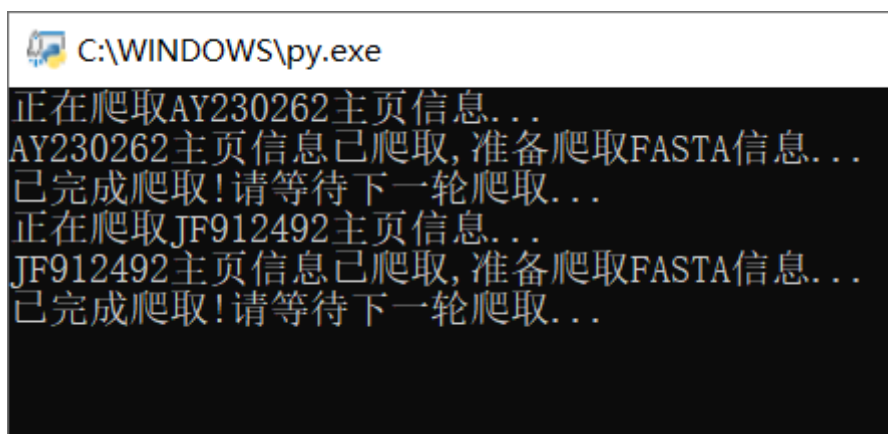


填写内容：

- 请求库：nucore
- 序列号：AY230262,JF912492
- 是
- 是
- 路径：C:/Users/cien/Desktop/download.txt
  - 需要修改成自己的路径！

点击“抓取”，提交。

在爬取过程中，Python对话框还会有提示信息输出：



```
C:\WINDOWS\py.exe
正在爬取AY230262主页信息...
AY230262主页信息已爬取, 准备爬取FASTA信息...
已完成爬取!请等待下一轮爬取...
正在爬取JF912492主页信息...
JF912492主页信息已爬取, 准备爬取FASTA信息...
已完成爬取!请等待下一轮爬取...
```

抓取结束后，我们还可以在自己的库中看到抓取下来的信息：

数据库				
添加新序列 删除序列 下载目标序列 清空数据库				
类型	注册号	名称	长度	来源
Nucleotide	AY230262	Homo sapiens NANOG m	918	Homo sapiens (human)
Nucleotide	JF912492	Heterocephalus glaber TR	2523	Heterocephalus glaber (n

同时，我们在download.txt文件中也找到了整合好的文件：

download.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

>AY230262.1 Homo sapiens NANOG mRNA, partial cds

```
ATGAGTGTGGATCCAGCTTGTCCTTGAAGCATCCGACTGTAAAGAATCTTCAC
CTATGCTGTGATTTGTGGGCTGAAGAAACTATCCATCCTTGCAAATGTCTTCTGCTGAGATGCCTCA
CACGGAGACTGTCTCTCTCTTCTTCTCCATGGATCTGCTTATTCAGGACAGCCCTGATTCTTCCACC
AGTCCCAAAGGCAAACAACCCACTTCTGCAGAGAAGAGTGTGCAAAAAAGGAAGACAAGGTCCCGGTCA
AGAAACAGAAGACCAGAACTGTGTTCTTCCACCCAGCTGTGTGTACTCAATGATAGATTTCAGAGACA
GAAATACCTCAGCTCCAGCAGATGCAAGAACTCTCCAACATCCTGAACCTCAGCTACAAACAGGTGAAG
ACCTGGTTCCAGAACCAAGAGAATGAAATCTAAGAGGTGGCAGAAAAACAACCTGGCCGAAGAATAGCAATG
GTGTGACGCAGAAGGCCTCAGCACCTACCTACCCAGCCTTACTCTTCTACCACCAGGGATGCCTGGT
GAACCCGACTGGGAACCTTCCAATGTGGAGCAACCAGACCTGGAACAATTCAACCTGGAGCAACCAGACC
CAGAACATCCAGTCTGGAGCAACCACTCCTGGAACACTCAGACCTGGTGACCCAATCCTGGAACAATC
AGGCCTGGAACAGTCCCTTCTATAACTGTGGAGAGGAATCTCTGCAGTCTGCATGCAGTTCAGCCAAA
TTCTCTGCCAGTGACTGGAGGCTGCCTTGGAGCTGCTGGGGAAGGCCTTAATGTAATACAGCAGACC
ACTAGGTATTTAGTACTCCACAAACCATGGATTTATTCCTAAACTACTCCATGAACATGCAACCTGAAG
ACGTGTGA
```

>JF912492.1 Heterocephalus glaber TRPV1 (Trpv1) mRNA, complete cds

```
ATGAAGAAATGGGCGAGTATAGACTCAAGGGAGTCTGAGCACCCACCCCAAGAGGAGGACTCCAGCCTGG
ACCCCCAGATACAGATGCTAACTCCAAGACACCTCCAGCCAAGCCCCATATTTTCTGTGAGCAAGAG
CCGTACCCGGCTCTTGGGAAAGGTGACTCGGAGGAGTTGTACCTATGGATTGCTTACGAGGAAGGA
GAACCAAGTTTCTGCCGACCATCACAGTCAGCTCTGTGGTCATCAGTCCAAGGCCTGGGACGGTCCCA
CCTGTGCCAGGCAGCTGTCCAGGACTCCATCCCTGCCAGTGCTGAAAAGCCACTCAAGCTCTATGATCG
GAGGAGCATCTTTGATGCTGTCGCTCAGAACAACTGCCAGGAAGTGGACAGCCTGCTGCCCTTCTGAAG
```

这样的文件就方便我们后续分析了！