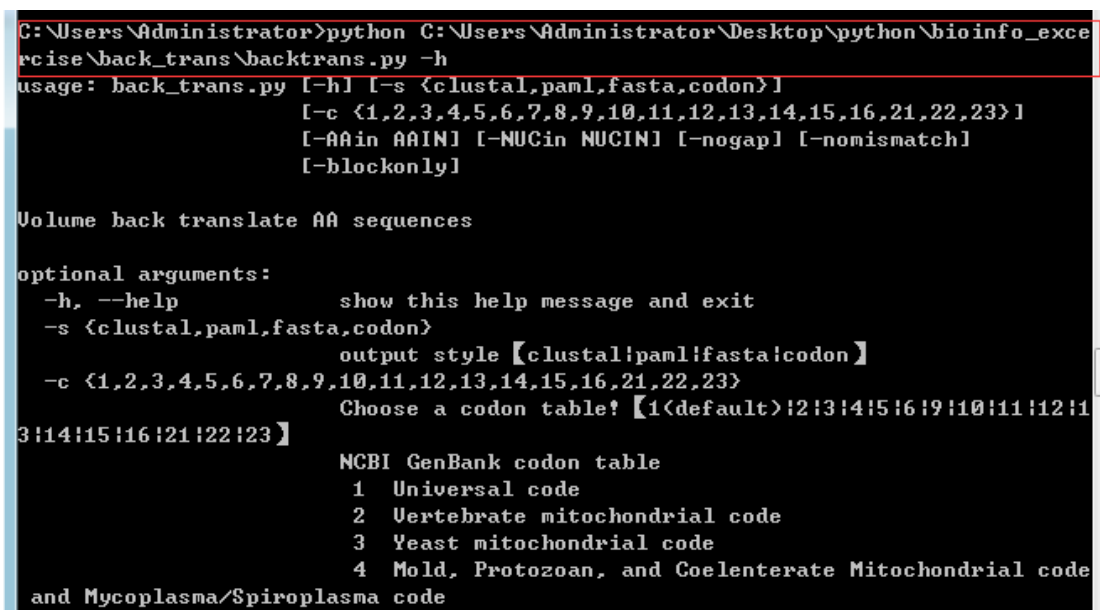


絮语：前一阵子建树的时候需要用到 codon 序列，但是一想到要在 MEGA 里面 codon 比对 13 个基因懒癌就犯了。于是到网上各种找合适的软件想写一个批运行脚本，clustal、muscle、prank、guidance 都跪了，最终找到一个 perl 脚本可以将 AA 序列回译为核苷酸序列，于是联想到可以先用 mafft 比对 AA 序列，然后回译为 codon，也相当于间接实现了 mafft 的 codon 比对【还支持输入 paml 格式】。

1.回归正题，运行需要 2 个文件，backtrans.py 和 pal2nal.pl；需要同时安装 python (3 及以上) 和 perl (5) 并配置环境变量。backtrans.py 脚本是基于 python3.43 编写的，可以支持批量运行、单文件运行（相当于比 pal2nal 多一个批运行功能）；pal2nal.pl 是 perl 5 编写，运行它需要先安装 perl5( <https://www.perl.org/get.html> 下载 strawberry perl，安装完即可运行)。

2. backtrans.py 有 8 个参数（所有参数均可选，未输入的参数均使用默认设置），-h 是查看帮助信息，如图：



```
C:\Users\Administrator>python C:\Users\Administrator\Desktop\python\bioinfo_exercise\back_trans\backtrans.py -h
usage: back_trans.py [-h] [-s {clustal,paml,fasta,codon}]
                    [-c {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,21,22,23}]
                    [-AAin AAin] [-NUCin NUCin] [-nogap] [-nomismatch]
                    [-blockonly]

Volume back translate AA sequences

optional arguments:
  -h, --help            show this help message and exit
  -s {clustal,paml,fasta,codon}
                        output style {clustal;paml;fasta;codon}
  -c {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,21,22,23}
                        Choose a codon table! {1<default>|2|3|4|5|6|9|10|11|12|13|14|15|16|21|22|23}

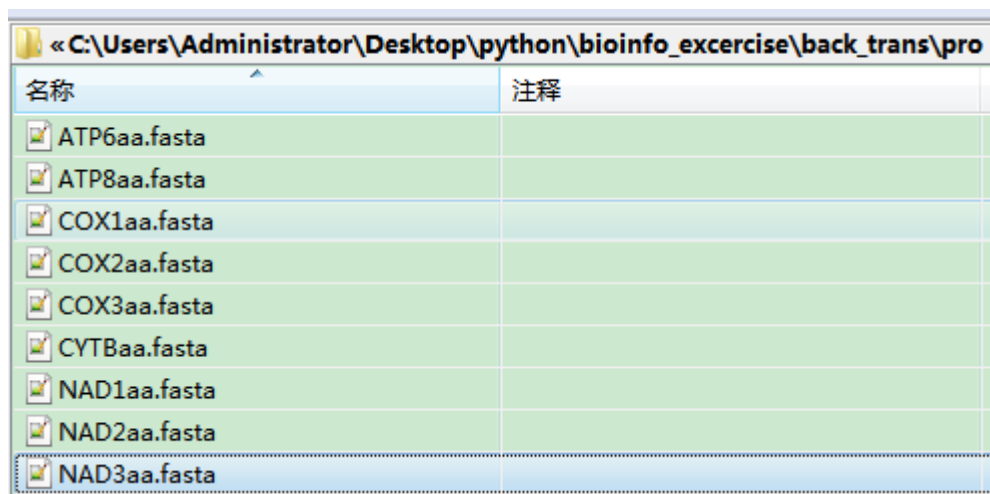
NCBI GenBank codon table
  1 Universal code
  2 Vertebrate mitochondrial code
  3 Yeast mitochondrial code
  4 Mold, Protozoan, and Coelenterate Mitochondrial code
  and Mycoplasma/Spiroplasma code
```

-s 是指定输出文件的格式，支持 4 种格式：clustal、paml、fasta 和 codon【默认 fasta 格式】

- c 是指定密码表，支持 19 种密码表（与 NCBI 密码表相对应）【默认通用密码表】
- AAin 是指定比对好的氨基酸序列（文件或文件夹均可）【默认脚本路径下的 pro 文件夹】
- NUCin 是指定氨基酸序列对应的核苷酸序列（文件或文件夹取决于氨基酸的输入）【同 AA】
- nogap 删除比对文件中有 gap 的一列【默认不删除】
- nomismatch 删除氨基酸序列与核苷酸序列之间不匹配的密码子【默认不删除】
- blockonly 只回译用户标记了 ‘#’ 位置的密码子（前提是要 clustal 格式），标记方法详见 test.aln 文件。【默认不使用】

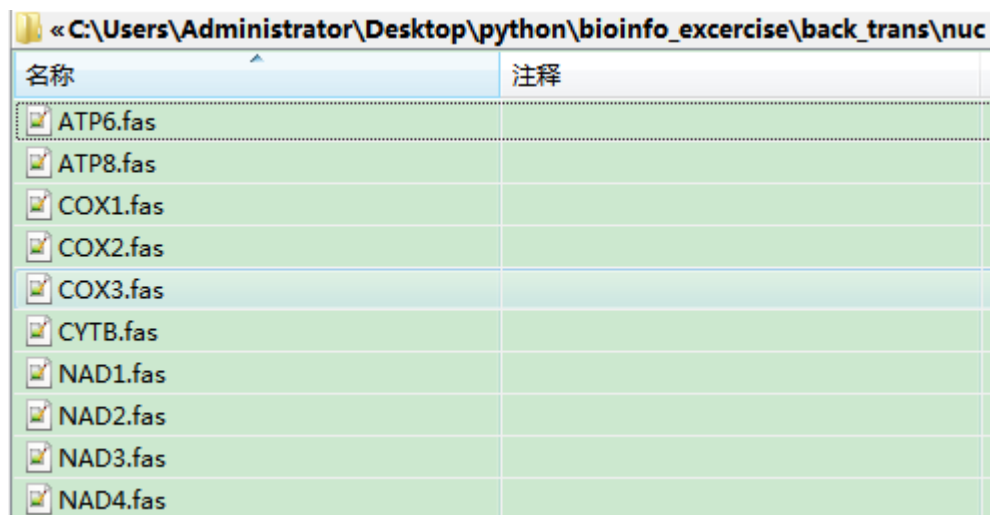
3.运用此脚本的前提是氨基酸和核苷酸文件的序列顺序必须一致；如果要批运行，那么输入的氨基酸与核苷酸文件夹里面的文件个数和文件顺序必须一样，如图：

这是 pro 文件夹里面的文件及顺序



名称	注释
ATP6aa.fasta	
ATP8aa.fasta	
COX1aa.fasta	
COX2aa.fasta	
COX3aa.fasta	
CYTBaa.fasta	
NAD1aa.fasta	
NAD2aa.fasta	
NAD3aa.fasta	

这是 nuc 文件夹里面的文件及顺序：



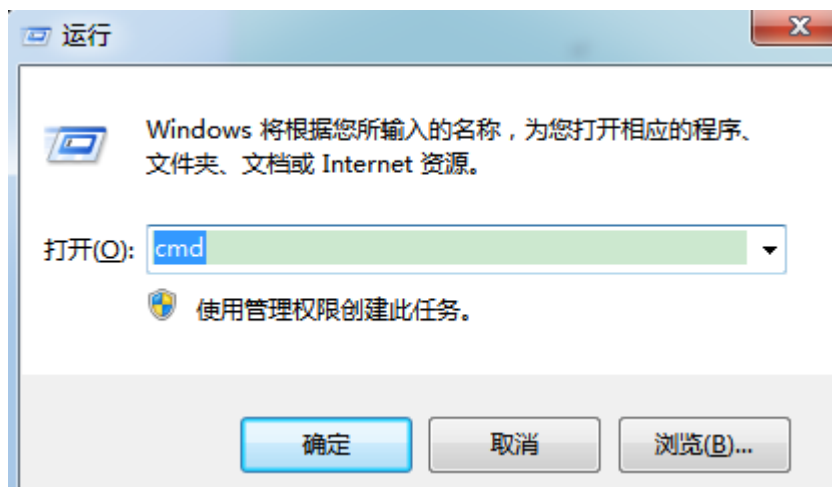
« C:\Users\Administrator\Desktop\python\bioinfo\_exercise\back\_trans\nuc

名称	注释
ATP6.fas	
ATP8.fas	
COX1.fas	
COX2.fas	
COX3.fas	
CYTB.fas	
NAD1.fas	
NAD2.fas	
NAD3.fas	
NAD4.fas	

注：nuc 文件夹里面的文件名被用于命名输出名。

运行示例：

首先打开运行，输入 cmd，出现 dos 命令框



1.文件夹形式批量运行①：氨基酸序和核苷酸序列的文件夹可以不在同一路径，运行时只需要将文件夹拖到对应的参数后面即可，如图（pro 文件夹在 mafft 文件夹内，而 nuc 文件夹存放于桌面）：

```
C:\Users\Administrator>python C:\Users\Administrator\Desktop\scripts\backtrans.py -AAin F:\software\mafft\mafft-win\pro -NUCin C:\Users\Administrator\Desktop\nuc -s fast a -c 2
```

2.文件夹形式批量运行②：当氨基酸和核苷酸序列所在文件夹与脚本所在文件夹相同时，输入时可以不用加上路径，如图：

nuc	文件夹	<文件夹>	2016/4/3 17:03:48	2016/4/3 10:02:14
pal2nal.v14	文件夹	<文件夹>	2016/4/7 21:38:56	2016/4/7 21:38:56
pro	文件夹	<文件夹>	2016/4/7 22:39:40	2016/4/4 20:55:58
backtrans.py	Python File	4,633	2016/4/7 22:17:16	2016/4/3 22:57:58

```
C:\Users\Administrator>python C:\Users\Administrator\Desktop\scripts\backtrans.py -AAin pro -NUCin nuc -s paml -c 2
```

3.单文件形式运行：当只有一对文件需要运行时，输入形式与上述类似（注意加上文件后缀）；这里我们可以尝试一下 test 数据的-nogap 和-blockonly 参数的效果。

注：这里 2 个 test 文件均存放于脚本所在文件夹

nuc	文件夹	<文件夹>	2016/4/3 17:03:48	2016/4/3 10:02:14
pal2nal.v14	文件夹	<文件夹>	2016/4/7 21:38:56	2016/4/7 21:38:56
pro	文件夹	<文件夹>	2016/4/7 22:39:40	2016/4/4 20:55:58
backtrans.py	Python File	4,633	2016/4/7 22:17:16	2016/4/3 22:57:58
guide.docx	Microsoft ...	0	2016/4/7 21:40:14	2016/4/7 21:40:14
pal2nal.pl	PL 文件	65,926	2011/12/2 13:14:32	2016/4/3 10:02:42
README.txt	文本文档	2,879	2016/4/7 22:44:28	2016/4/4 22:01:34
test.aln	ALN 文件	719	2011/12/2 13:09:22	2016/4/5 10:50:28
test.nuc	NUC 文件	1,768	2011/12/2 13:09:22	2016/4/5 10:50:28

```
C:\Users\Administrator>python C:\Users\Administrator\Desktop\scripts\backtrans.py -AAin test.aln -NUCin test.nuc -s clustal -c 2 -nogap -blockonly
```

所有方式运行完的结果均保存在脚本所在文件夹的 output 文件夹里：

nuc	文件夹	<文件夹>	2016/4/3 17:03:48	2016/4/3 10:02:14
output	文件夹	<文件夹>	2016/4/7 23:01:54	2016/4/7 23:01:54
pal2nal.v14	文件夹	<文件夹>	2016/4/7 21:38:56	2016/4/7 21:38:56
pro	文件夹	<文件夹>	2016/4/7 22:39:40	2016/4/4 20:55:58
~\$dance.docx	Microsoft ...	162	2016/4/7 23:02:50	2016/4/7 23:02:50
back_trans.py	Python File	2,653	2016/4/3 20:32:48	2016/4/3 20:32:48
backtrans.py	Python File	4,633	2016/4/7 22:17:16	2016/4/3 22:57:58