# 运行记录

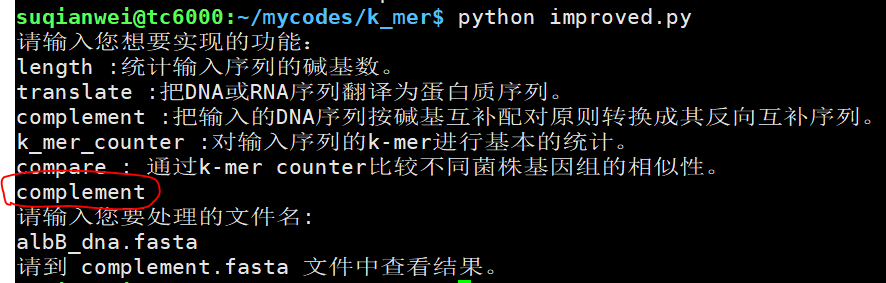
苏茜薇

记录的比较乱，图比较多，请打开视图-显示-导航窗格来查看，方便跳转。

## 实现complement和translate的功能

测试这两个功能的DNA序列为albB\_dna.fasta, 它是我本科的时候用过的一个蛋白序列，这里提供的是它的DNA序列，需要先把该序列转变为反向互补序列，再对得到的反向互补序列进行翻译，才能得到蛋白序列。

### Complement功能





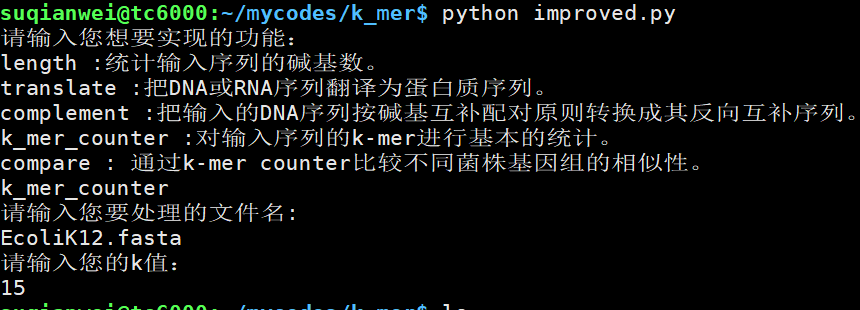
### Translate功能

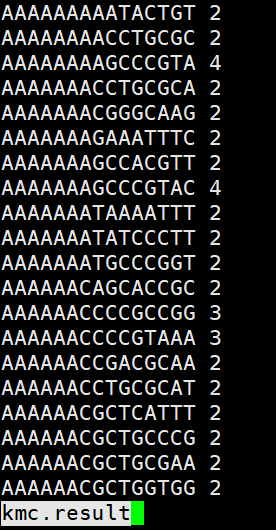


## k-mer counter 功能

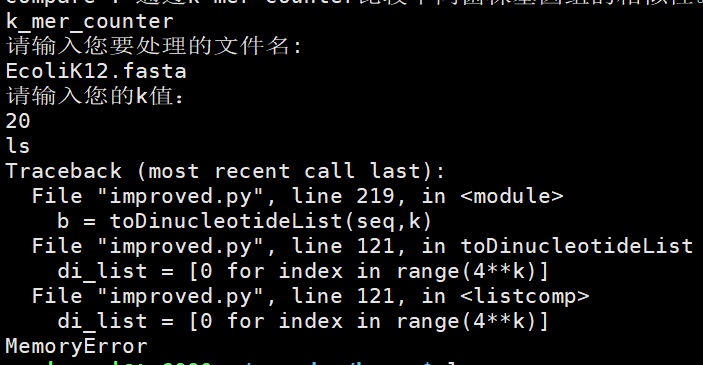
### 单个序列的k-mer计数

对于单个序列的k-mer 计数，使用**大肠杆菌的测试数据**，k=15时可以得到结果，k=20时就有内存错误。

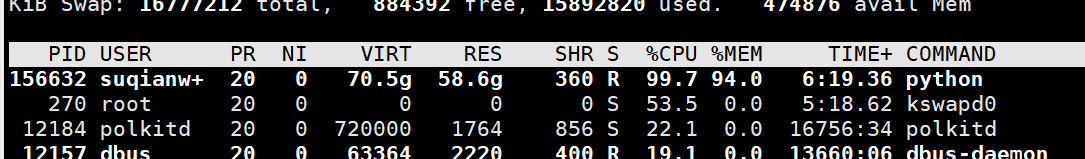




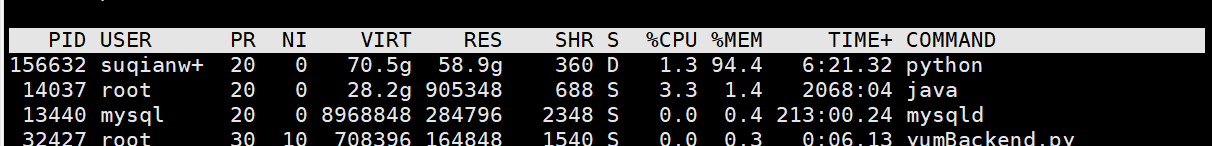
K=20时出现内存错误



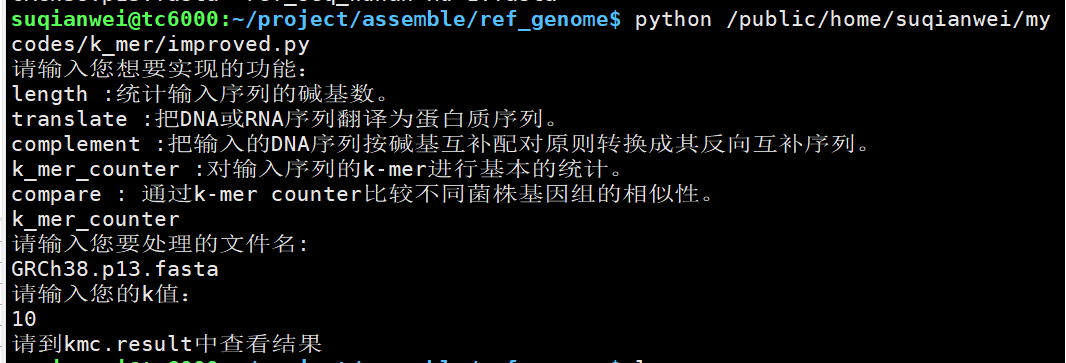
开始时cpu和内存占用都很高



一段时间后，cpu占用降下来了，但内存占用持续，再过一会儿就报内存错误

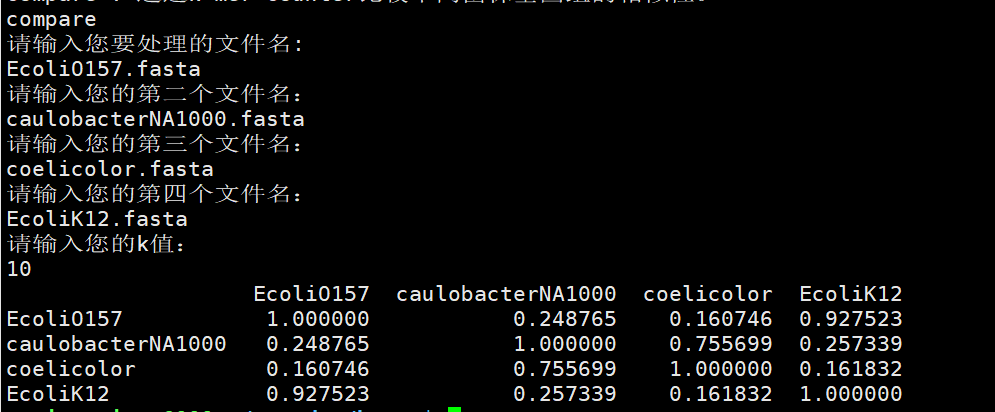


对**人的基因组序列**GRCh38.p13.fasta进行的k-mer 计数，k=10时，几分钟内运行完毕。我没有尝试k更大时的情况。另外，这个基因组有很多N，2个M，1个R，我直接把这些奇怪的字符去掉了，再来做的k-mer 计数。



### 通过k-mer对四个序列进行相似性比较

K值设置为10时，程序可以较快给出结果（几分钟内）



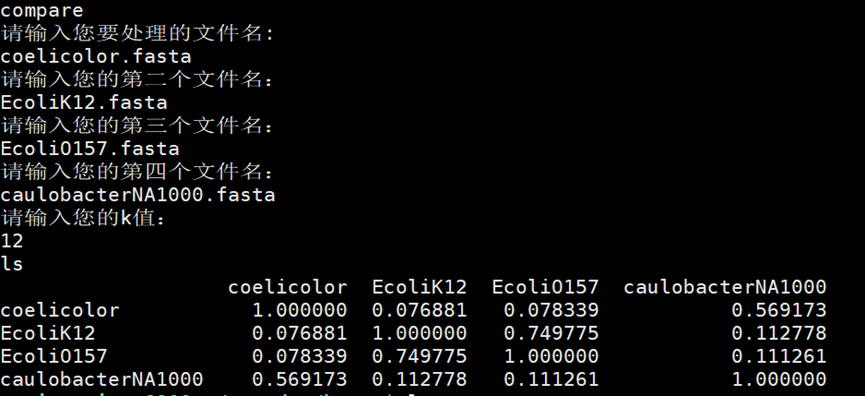
分析一下这个结果，两个大肠杆菌很相似，相似性有0.93；caulobacterNA1000与coelicolor较相似，相似性有0.75，coelicolor是我本科做过的菌，天蓝色链霉菌，是一个格兰氏阳性细菌。

这个相关性矩阵可以做个可视化，等我去R里面搞个热图过来。

K值设置为15时，程序无法正常运行出结果：



K=12时，也还可以。

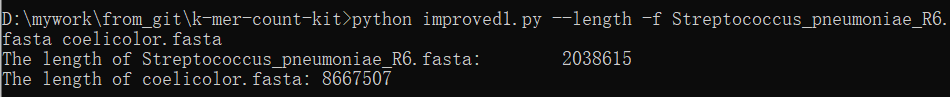


# 11月更新

1. 采用argparse来进行互动，极大的方便了程序的运行。
2. 对可处理文件的数目做了优化，更新后可以处理多个文件。

2.1 计算序列长度的功能(--length)

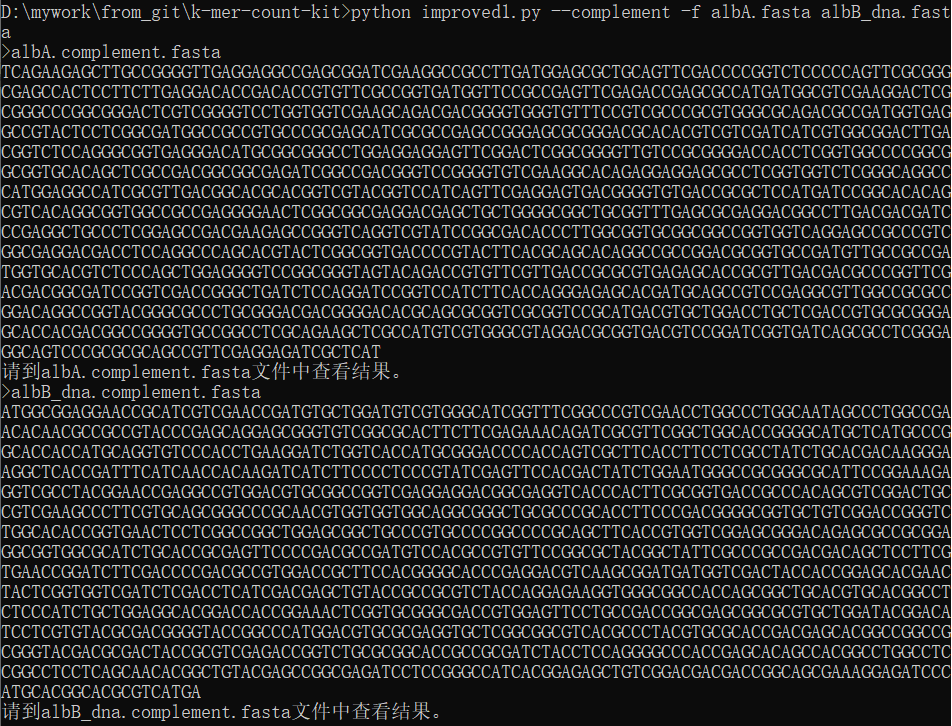
可以同时统计输入的多个文件的序列长度。



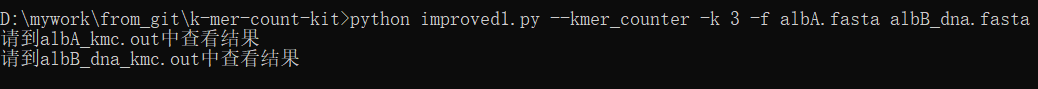
2.2 把DNA或蛋白质序列翻译为蛋白质序列的功能(--translate)



2.3 得到反向互补序列的功能(--complement)



2.4 kmer-counter的功能(--kmer\_counter)



2.5 计算不同序列之间的相关性(--compare)

现在可以对多个序列文件计算其相关性，不再局限于四个输入文件。

