

# 生信编程直播课程优秀学员学习心得及作业展示3

Original 王康利 生信技能树 2017-03-17 05:13

收录于合集

#学徒作业

117个

大家好，本次分享一下我在做**生信技能树——生信编程直播**第一题和第二题的一些方法和体会，最后附上自己学习之后的代码成果。

## 学习感悟

首先说明一下，我不算是完全从0开始学习，因为生物的知识python的语言之前都知道一点，但说实话，我的python距离真正的实践还差的很远，也没有常用所以基本忘完。

真的很感谢群主和老师们搭建这样一个学习的平台，并且讲解的内容会涉及到方方面面，从数据的下载讲解到编程的技巧思维，还有生物的知识等等。

所以如果你是刚开始学习，不用担心，当你做了1-2道题之后，你就会发现在老师们的引导下，你可以很快的入门。

如果你对于编程是完全的新手，不用担心，所有的老师都会告诉你如果遇到不懂的，直接先去网上搜索一下。

例如还没有装python，你可以直接去网上搜索一下“python下载”就可以下载安装python，如果你是跟着老师学习，老师同样给你推荐很好的一些编辑器使用。

如果遇到不懂的命令或者错误提示信息，大部分同样可以在网上找到答案或者在论坛跟别人讨论。当然如果你刚开始接触python，我会建议你先学习一些python的基本命令和语法，因为老师们不会讲解最基础的语法，还有你要使用一些非python自带的包，就要学会安装，pip基本可以搞定。你在刚开始学习的时候会发现很多不理解的地方，我的建议是先尝试去运行老师演示的代码，看看会发生什么，这样你就会慢慢的有了对编程的理解。

---

## 关于题目

下面说说和题目相关的内容，前两道题其实有些类似，我会从我的做题思路出发串讲一下。

更详细的内容，大家可以去我们的网站查看，有详细的问题解析和大家的做题情况及讨论。地址：<http://www.biotrainee.com/thread-625-1-1.html>

对于每道题，我一般会在看老师视频之前尝试自己做一遍，即使做的不对甚至做不出来，那么在看老师视频的时候就可以看到差距从而可以查漏补缺。

第一道题是探究人类外显子区域的长度，第二道题是hg19基因组序列的探究。

我跟东野老师学到了一个习惯：**面对一个问题不是马上去写代码，而是要先把问题解析清楚**，例如数据类型、问题说明，数据结构等分析出来，老手们的思路可能很快就把这些想清楚，但是新手最好还是先把这些想清楚，养成好习惯。

这两道题首先涉及到数据下载，感谢群主已经在问题描述里把链接都给出来了，节省我们学员的时间。接下来你需要了解你下载下来的数据结构以及你需要提取的数据信息，例如第一题需要的就是外显子的起始点和终止点着两列。

然后是你想统计的结果和你要储存的数据结构，例如第一题就是把每个exon的起始到终止存在字典里，把它的长度累加起来。最后就是输出结果了，可以直接打印在屏幕或者输入到一个文件里。前面我提到了每次我都会先自己做一次，然后看老师视频，这样就会学到新知识并且比较深刻，例如第一题我自己做的时候就没有考虑外显子的重复，对每个外显子都统计了，最后跟老师答案比较就发现自己没有去除那些完全重复的外显子，另外对比代码也可以学到新的包和高级快速的技巧。

最后提醒大家，**一定要对自己的代码写上清楚的注释信息**，这样以后自己再看或者跟别人交流就会很清楚，这一点我之前做的不太好也没重视，现在很多时候就要重新理解浪费时间。

这里讲解了我的一些思路和体会，论坛有更多人的分享需要你们去学习探索。做了这两道题之后，再写代码的时候你会发现自己多了很多技巧，例如你的代码结构会更成熟，思维更清晰了，包的使用也会更熟练。

---

## 关于老师

如果你学的是python，做完前两道题，你就会接触到群主和python的3个讲师了。

- 群主会把问题给你分析的很清楚，包括数据准备，演示代码和结果部分，这都是很细心和用心的。

- 东老师会给你推荐一些python的工具技巧，而且他的代码我认为是最直接和最精炼的答案，他的编程思维也是和我这样的入门者的编程思维最相近的老师。
- 李治鑫老师会首先给你讲解针对问题的最基本的命令，然后会教你使用python包来处理类似问题使你更加得心应手，包的使用在python里是非常重要的技能了。
- 东野老师是计算机出身，他的编程思维是非常规范和严谨的，同时他在第二讲的时候从初学者的起点给我们示范了如果用编程处理生信问题，从分析问题到一步步解决问题再到优化代码，讲解非常详细。

三位老师都非常优秀又风格不同，你可以从他们的演示学到很多编程思维和技巧，慢慢就会养成自己的编程习惯了。几位老师准备视频都付出了很多时间，都把自己的技巧无保留的给大家讲解了。如果你还没有开始，马上加入我们开始学习生信编程吧。

---

## 题目代码和注释

最后附上自己的代码成果和大家交流，也算是给老师交作业，从一个python生信编程的文盲到入门，感谢老师们的讲解付出。

第一题需要处理的数据类型算是普通文本文件：

```
#chromosome  nc_accession  gene      gene_id  ccds_id  ccds_status  cds_strand  cds_from  cds_to  cds_locations  match_type
1            NC_000001.8    LINC00115  79854    CCDS1.1  Withdrawn    -          -          -      None
1            NC_000001.11   SAMD11    148398   CCDS2.2  Public      +          925941  944152  [925941-926012, 930154-930335, 931038-931088, 935771-935895,
1            NC_000001.11   RUC2L     26159    CCDS3.1  Public      -          944693  959239  [944693-944799, 945056-945145, 945517-945652, 946172-946285,
```

第一题的代码及我的注释：




```

##第二题：探索基因组序列
##学习李老师的代码
##方案一，思路简单直接，但是如果输入基因组的话运行时间超长，供初学者学习，不推荐
#导入需要的包，套路
import os
from collections import OrderedDict
#修改路径，套路
os.chdir("./")
#初始化变量，套路
chr_dict = OrderedDict()
temp_chr = ""
#读文件，套路
with open("UCSC_hg19_chrAll.fasta","rt") as f:
    for line in f:
        line = line.strip()
        if line.startswith(">"): #判断首字符，如果是>代表一个新序列的注释行，即染色体信息
            temp_chr = line
            chr_dict[temp_chr] = "" #在字典里添加这个染色体为一个元素
        else:
            chr_dict[temp_chr] += line #如果是序列信息就把它累加在字典的该染色体上

for seqName,seq in chr_dict.items(): #最后输出依次输出每个染色体的碱基统计信息
    A = seq.count("A") + seq.count("a") #注意基因组会有小写碱基
    T = seq.count("T") + seq.count("t")
    C = seq.count("C") + seq.count("c")
    G = seq.count("G") + seq.count("g")
    N = seq.count("N") + seq.count("n")
    print(seqName,A,T,C,G,N)


```

 生信技能树

```

##方案二，使用pysam包，推荐，非常快速。
#导入需要的包，套路
import pysam #这个包李老师推荐的，处理组学的数据类型会很方便哦
import os
#修改路径，套路
os.chdir("./")
#用pysam包处理fasta文件，它会建立索引，再次使用时就会很快速方便
hg19 = pysam.FastaFile("UCSC_hg19_chrAll.fasta")
dir(hg19) #查看列表可使用的方法
list(zip(hg19.references,hg19.lengths)) #把染色体和其对应的序列长度组合在一起输出
for seqName in hg19.references: #依次输出每条染色体的碱基统计信息
    seq = hg19.fetch(seqName)
    A = seq.count("A") + seq.count("a") #注意会有小写碱基
    T = seq.count("T") + seq.count("t")
    C = seq.count("C") + seq.count("c")
    G = seq.count("G") + seq.count("g")
    N = seq.count("N") + seq.count("n")
    print(seqName,A,T,C,G,N)
#pysam包非常有用，但是我还没有深入学习，大家有兴趣可以继续学习它。
#另外基因组有很多值得探索的信息，论坛上大家的作业和讨论非常值得大家去继续学习。

```

 生信技能树

本文编辑：思考问题的熊





收录于合集 #学徒作业 117

上一篇

生信编程直播课程优秀学员作业展示2

下一篇

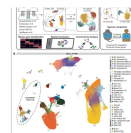
生信编程直播课程优秀学员作业展示1

People who liked this content also liked

EDCI, HATU, CDI等酸胺缩合试剂副产物和中间态结构和MS分析  
有机合成路线



人肺内皮细胞整合单细胞图谱  
生信菜鸟团



TLC点板，产物极性忽大忽小，咋回事！  
有机合成路线

