python3生信入门-基因组fasta基本信息统计（序列数、总长度、N50、L50等）

测试fasta文件（无提取码）：<https://pan.baidu.com/s/1yYQNxty5hBdcpoA8xkrQUg>

示例脚本（提取码 lpsq）：https://pan.baidu.com/s/14zifcZWq8oZqqHA4-HNg\_g

## 说明

网盘中提供的测试fasta文件“Bacillus\_subtilis/4-assembly/Bacillus\_subtilis.scaffolds.fasta”，为某细菌的基因组组装草图（组装到scaffolds水平，每条序列为一条scaffold，scaffold命名为scaffold\_1、scaffold\_2、scaffold\_3……）。

现在我们需要使用python3，对该基因组fasta文件的基本信息进行统计，包括：

（1）该基因组所含scaffolds总数、碱基总个数、平均GC含量百分比、N碱基个数；

（2）scaffolds平均长度、最长scaffolds、最短scaffolds，N50、N90、L50、L90值；

（3）长度大小分别在1000000bp、100000 bp、10000bp、1000bp以上的scaffolds的总个数及长度。

并最终输出一个统计结果文件，请根据要求编写一个简单的py脚本。

## 示例

一个示例脚本如下（可参见网盘附件“fasta\_stat.1.py”，请使用python3.6及以上版本测试）。

打开fasta文件“Bacillus\_subtilis.scaffolds.fasta”，使用循环逐行读取其中的序列id及碱基组成信息，并分别统计各序列的所有碱基总个数、GC碱基总个数以及N碱基总数（G、C、N等字符可使用正则表达式识别，re模块的re.findall命令）；将序列id及该序列的长度数值以字典的形式存储（字典样式{'id':'length'}），同时统计该fasta文件中的总碱基数、GC碱基总数以及N碱基总数。

对字典进行排序后（按值排序，即按scaffolds的长度大小由大到小排序，可使用lambda命令完成），统计L50、N50、L90、N90等指标，以及按scaffolds长度分布统计个数及总长度。

最后打开新文件“fasta.stat.txt”，将各统计值写入其中。

#!/usr/bin/env python3.6

# -\*- coding: utf-8 -\*-

#导入模块，初始传递命令、变量等

import re

input = 'Bacillus\_subtilis.scaffolds.fasta'

output = 'fasta.stat.txt'

dict = {}; dict\_refer = {}

seqs\_sum = 0

base\_sum = 0

GC\_sum = 0

N\_sum = 0

Ln = 0; base\_sum\_n = 0

N50 = 0; N75 = 0; N90 = 0

n1000000 = 0; n1000000\_len = 0

n100000 = 0; n100000\_len = 0

n10000 = 0; n10000\_len = 0

n1000 = 0; n1000\_len = 0

#读入 fasta 文件，统计 scaffolds（每条 + 所有）总数、长度、GC 含量等基本信息

with open(input, 'r') as read\_fas:

for line in read\_fas:

if line[0] == '>':

key = line.strip('[ >\n]')

dict[key] = 0

seqs\_sum += 1

else:

value = line.strip()

seqs\_len = len(value)

base\_sum += seqs\_len

dict[key] += seqs\_len

GC\_sum += len(re.findall('[gcGC]', value))

N\_sum += len(re.findall('[nN]', value))

read\_fas.close()

#统计 N50、N75、N90 以及各长度区间 scaffolds 数量及长度

dict\_sort = sorted(dict.items(), key = lambda x:x[1], reverse = True)

base\_sum\_n50 = 5 \* base\_sum/10

base\_sum\_n75 = 7.5 \* base\_sum/10

base\_sum\_n90 = 9 \* base\_sum/10

for value in dict\_sort:

base\_sum\_n += value[1]

Ln += 1

if N50:

if N75:

if base\_sum\_n >= base\_sum\_n90:

N90 = value[1]

L90 = Ln

break

else:

if base\_sum\_n >= base\_sum\_n75:

N75 = value[1]

L75 = Ln

else:

if base\_sum\_n >= base\_sum\_n50:

N50 = value[1]

L50 = Ln

for value in dict\_sort:

if value[1] >= 1000000:

n1000000 += 1

n1000000\_len += value[1]

if value[1] >= 100000:

n100000 += 1

n100000\_len += value[1]

if value[1] >= 10000:

n10000 += 1

n10000\_len += value[1]

if value[1] >= 1000:

n1000 += 1

n1000\_len += value[1]

if value[1] < 1000:

break

#统计总览，并输出结果

basic\_stat = open(output, 'w')

print(f'Total scaffolds\t{seqs\_sum}', file = basic\_stat)

print(f'Total base (bp)\t{base\_sum}', file = basic\_stat)

print(f'Total N (bp)\t{N\_sum}', file = basic\_stat)

print(f'Average length (bp)\t{int(base\_sum/seqs\_sum)}', file = basic\_stat)

print(f'Longest scaffold (bp)\t{dict\_sort[0][1]}', file = basic\_stat)

print(f'Shortest scaffold (bp)\t{dict\_sort[-1][1]}', file = basic\_stat)

print(f'L50\t{L50}', file = basic\_stat)

print(f'N50\t{N50}', file = basic\_stat)

print(f'L75\t{L75}', file = basic\_stat)

print(f'N75\t{N75}', file = basic\_stat)

print(f'L90\t{L90}', file = basic\_stat)

print(f'N90\t{N90}', file = basic\_stat)

print(f'GC (%)\t{round(100 \* GC\_sum / base\_sum, 2)}', file = basic\_stat)

print(f'Total scaffolds (>= 1000000 bp)\t{n1000000}', file = basic\_stat)

print(f'Total scaffolds (>= 100000 bp)\t{n100000}', file = basic\_stat)

print(f'Total scaffolds (>= 10000 bp)\t{n10000}', file = basic\_stat)

print(f'Total scaffolds (>= 1000 bp)\t{n1000}', file = basic\_stat)

print(f'Total length (>= 1000000 bp)\t{n1000000\_len}', file = basic\_stat)

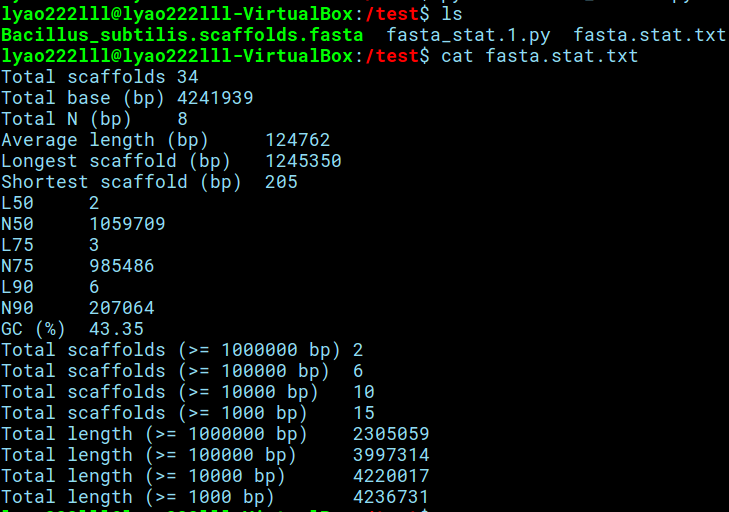
print(f'Total length (>= 100000 bp)\t{n100000\_len}', file = basic\_stat)

print(f'Total length (>= 10000 bp)\t{n10000\_len}', file = basic\_stat)

print(f'Total length (>= 1000 bp)\t{n1000\_len}', file = basic\_stat)

basic\_stat.close()

输出结果文件“fasta.stat.txt”其内容如下。



其中：

Total scaffolds：scaffolds总数；

Total base (bp)：碱基总数；

Total N (bp)：N碱基个数；

Average length (bp)、Longest scaffold (bp)、Shortest scaffold (bp)：scaffolds平均长度、最长scaffold以及最短scaffold；

L50、L75、L90、N50、N75、N90：基因组L50、N50等指标；

GC (%)：基因组平均GC含量百分比

Total scaffolds (>= 1000000 bp)、(>= 100000 bp)、(>= 10000 bp)、(>= 1000 bp)：长度分布大于1000000bp、100000bp、10000bp、1000bp的scaffolds个数；

Total length (>= 1000000 bp)、(>= 100000 bp)、(>= 10000 bp)、(>= 1000 bp)：长度分布大于1000000bp、100000bp、10000bp、1000bp的scaffolds总长度。

补充：

网盘附件“fasta\_stat.py”为添加了命令传递行的python3脚本，可在shell中直接进行目标文件的I/O处理。指定读取的fasta文件，以及输出文件路径和名称即可。

使用其统计上文测试fasta文件中所含序列的基本信息的方式如下所示。

#python3 fasta\_stat.py -h

python3 fasta\_stat.py -i Bacillus\_subtilis.scaffolds.fasta -o fasta.stat.txt

