# 计算生物学第二次作业

## 一、实验分析：

本次实验使用了Python语言编写了隐马尔可夫模型Viterbi算法，实现了在已知初始概率、输出概率、转移概率的情况下，通过测序数据推断序列真实的基因型。该算法的原理如下：

首先定义了以下参数：

# 状态的样本空间，0表示ZZ、1表示MM、2表示MZ

states = ['0','1','2']

# 初始概率，取对数是为了将后面乘算变为加算，从而简化计算。

S = [log(0.4), log(0.4), log(0.2)]

# 状态转移概率

T = [

[log(0.99\*0.99), log(0.01\*0.01), log(2\*0.99\*0.01)],

[log(0.01\*0.01), log(0.99\*0.99), log(2\*0.99\*0.01)],

[log(0.01\*0.99), log(0.01\*0.99), log(0.01\*0.01+0.99\*0.99)]

]

# 输出概率

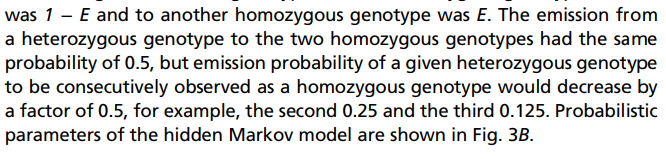
E = [

[log(0.97), log(0.03)],

[log(0.03), log(0.97)],

[log(0.25), log(0.25)]

]

这里输出概率第3行MZ生成MM、ZZ取0.25是因为以下原因：

大致意思就是，因为杂合MZ本来还有0.5概率生成杂合MZ，但观测值只有纯合MM、ZZ，如果纯合都取0.5的话，那么杂合观测为纯合的概率会每轮降低0.5倍，第一个为0.5、第二个为0.25、第三个为0.125。因此这里概率取0.5/2=0.25。

通过obs=input（）输入观测序列，由于python input输入的是字符串，需要通过replace = {'0': 0, '1': 1}将输入的字符串转换为整数。n=len(obs)

获取序列的长度后，构建s = zeros((3, n))的3\*m矩阵来存放计算数据，构建path = zeros((3, n, 3))矩阵来记录每个数据的方向下标0处存储从第一行ZZ来的方向，下标1处存储从第二行MM来的方向，下标2处存储从第三行MZ来的方向，初始值均为0，若存在从某方向上来则将其对应下标的值置为1。构建数组LastPath=[]记录最终打分路径。

之后创建得分矩阵：

S(i,0)=S[i] + E[i][replace[obs[0]]]

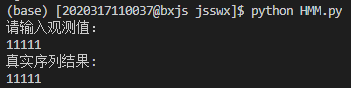
S(i,j)**=**

通过path来记录s（i，j）的方向，0表示从第一行ZZ来，1表示从第二行MM来，2表示从第三行MZ来。若上面一式最大path[i,j,0]=1，二式最大path[i,j,1]=1，三式最大path[i,j,2]=1。

最后通过回溯寻找最佳路径，首先判断最后一列的最大值，j == n-1，s[i][j] == max(s[0][j], s[1][j], s[2][j])时，通过x记录此时的i，并LastPath. append(x)将其存入数组LastPath中。之后通过前面path[x,j,k]==1记录的路径进行回溯，k值即为方向，x = k，LastPath.append(x)其存入数组LastPath中，相应的数值代表了相应的基因型。由于通过回溯记录的序列是反的，因此需要LastPath.reverse()将数组反转，最后print输出结果。

## 二、运行结果：

① 验证程序的正确性，先输入5个基因的短序列：11111



验证该结果：下表可以看出结果正确

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 1 | 0 | 0 | 1 |
| 0 | -4.42 | -7.94 | -11.47 | -13.76 | -13.81 |
| 1 | -0.94 | -0.99 | -1.04 | -1.09 | -1.14 |
| 2 | -2.99 | -4.40 | -5.80 | -6.35 | -6.40 |

② 输入测序观测值

