

生物信息学概论 作业 1

算法思路: Needleman 和 Wunsch 算法

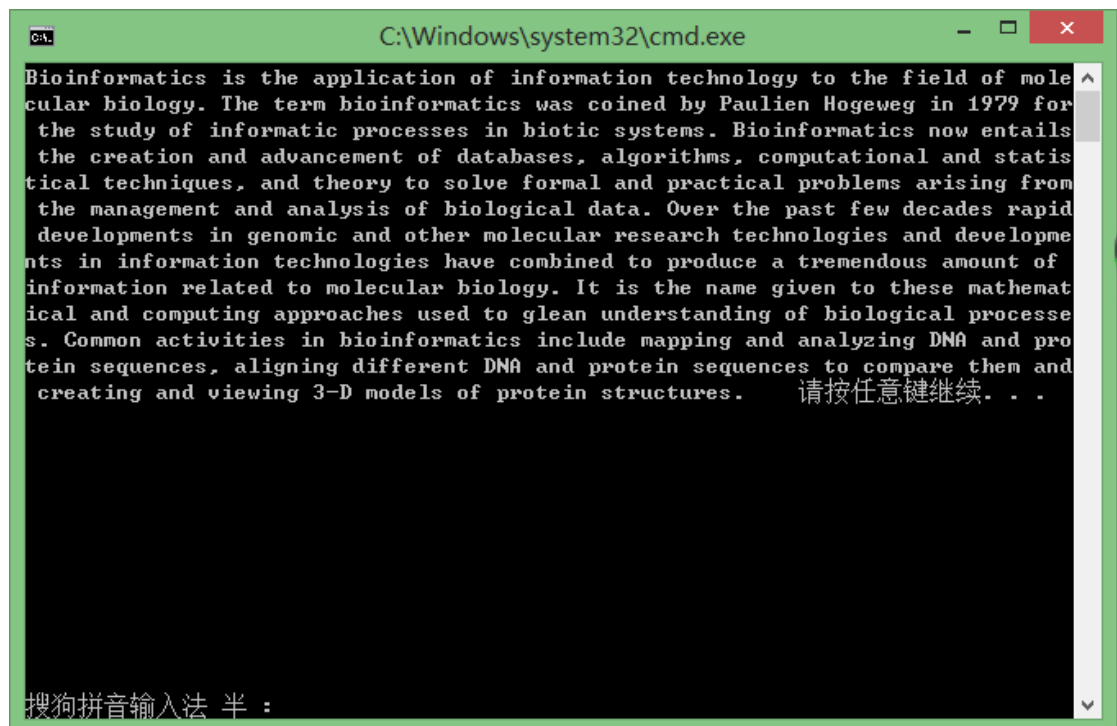
对于双序列比对, 构造得分矩阵, 通过给匹配、不匹配、空位匹配三种情形按照动态规划的方法, 以取正上方、正左方、左上角三个方向的邻居迭代至当前位置所打分数的最大值这个原则构造出每种序列匹配情况下的得分; 于此同时构造回溯矩阵, 回溯矩阵中每个位置记录对应回溯时当前位置回溯的方向, 然后采用回溯的方法, 从得分矩阵右下角开始回溯, 直至左上角, 回溯形成的路径就是双序列的最佳比对。

C++版:

主函数中读入序列信息 (LongestCommonSeq.txt 存储路径如右边, 请修改代码中相应部分), 给定匹配、不匹配、空位匹配的打分值。

全局序列比对函数: `char * Sequence_check(string sql, string sq2, int alpha0, int beta0, int gamma0);` 返回最大公共子串, 然后主函数输出。

输出结果:



```
C:\Windows\system32\cmd.exe
Bioinformatics is the application of information technology to the field of molecular biology. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures. 请按任意键继续. . .
搜狗拼音输入法 半 :
```

代码执行过程中, 我们能够明显感觉到 C++ 语言的速度优势, 整个程序在我的电脑上运行时间不到 10s.

R 语言版:

在一开始使用最单纯的 R 语言编写本次作业时，发现 R 的速度真是慢到无以复加。咨询了助教师兄的建议后，采用一种能结合 R 和 C++的方法 Rcpp 来编写本次作业。

Local_alignment.R 是主程序，读入序列信息 (LongestCommonSeq.txt 存储路径如右边，请修改代码中相应部分)，给定匹配、不匹配、空位匹配的打分值。

Sequense.cpp 编写函数 string sequense(string sql, string sq2, int alpha0, int beta0, int gamma0)，返回公共子串至主程序输出。

用 Rstudio 打开两个文件, 在工作区输入: `source('A:/学习/大四上/生物信息学概论/作业/Homework1/R/Local_alignment.R', encoding = 'UTF-8')` 打开文件，输出如下：

```
> source('A:/学习/大四上/生物信息学概论/作业/Homework1/R/Local_alignment.R', encoding = 'UTF-8')
[1] "Bioinformatics is the application of information technology to the field of molecular biology. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures."
```

对比 C++，采用基本 R 语言编写本次作业由于存在双循环，使得程序的运行速度非常慢；采用了结合 C++的 Rcpp 后，使得 R 利用了 C++的速度优势，整个程序的执行时间降到了与 C++版程序差不多的时间。

Python 版:

Local_alignment.py 是主程序，读入序列信息 (LongestCommonSeq.txt 存储路径如右边，请修改代码中相应部分)，给定匹配、不匹配、空位匹配的打分值。

Sequense_check.py 编写函数 `def sequense(sql, sq2, alpha0, beta0, gamma0)`，返回公共子串至主程序输出。

打开 cmd

输入 python

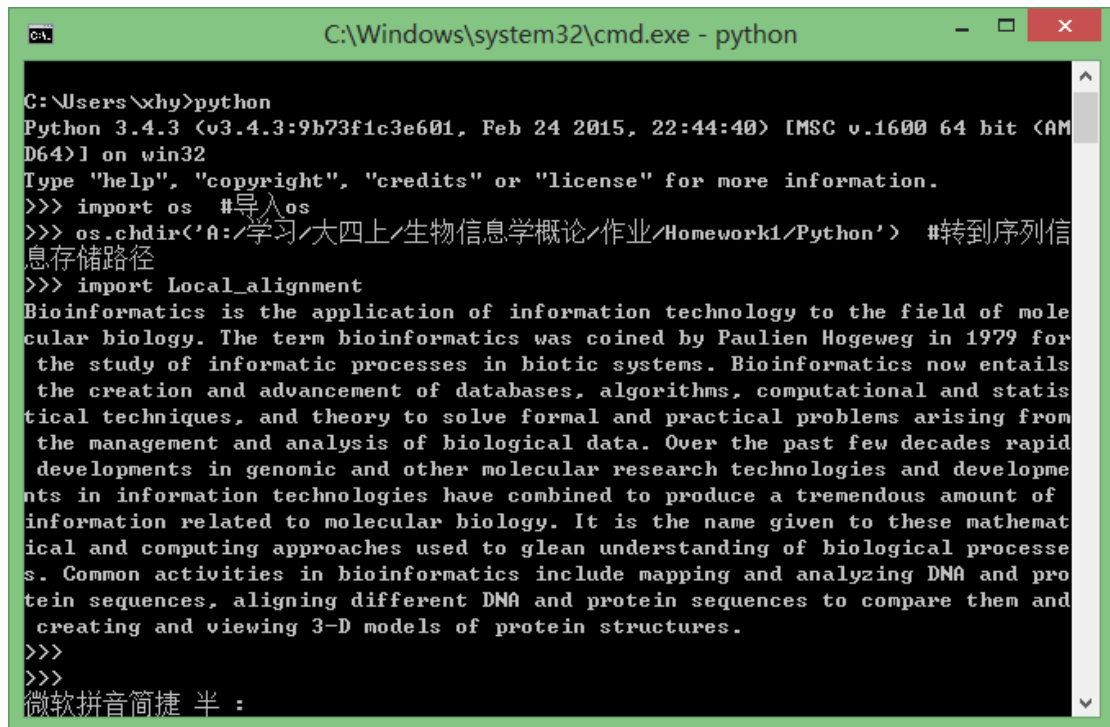
输入 `import os` #导入 os

`os.chdir('A:/学习/大四上/生物信息学概论/作业`

/Homework1/Python') #转到 Local_alignment.py 存储路径

import Local_alignment #运行主程序，如必要修改主程序中序列信息的存储路径

输出如下：



```
C:\Windows\system32\cmd.exe - python

C:\Users\xy>python
Python 3.4.3 (v3.4.3:9b73f1c3e601, Feb 24 2015, 22:44:40) [MSC v.1600 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import os #导入os
>>> os.chdir('A:/学习/大四上/生物信息学概论/作业/Homework1/Python') #转到序列信息存储路径
>>> import Local_alignment
Bioinformatics is the application of information technology to the field of molecular biology. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures.
>>>
>>>
微软拼音简捷 半 :
```

Python 和 R 语言都属于动态语言，在没做优化的情况下，运行速度都特别慢。Python 版在我的电脑上运行了 10min 左右才出结果，我猜想这可能跟我的算法和使用的 IDE 也有关，可能换一个不同的 IDE 和算法能提高 Python 的运行速度。