|  |
| --- |
|  |
| 蛋白质序列筛选程序说明 |
|  |
|  |
| **李启明** |
| **2019-05-18** |

|  |
| --- |
| 本文档编写目的为说明本程序的运行环境和设计思路及使用方法 |

编程语言：Python

运行环境：Python3以上

依赖库： Biopy

* Part A:初步筛选程序

以下使用字母 P T S 代表 氨基酸 氨基酸 氨基酸

因（此处补充粘液素PTS含量高的原因），所以首先要从数据库中筛选出符合粘液素蛋白特征的序列。

**处理逻辑：**取得程序所在路径下所有fasta格式数据库，遍历每数据库中序列，如果序列中PTS含量满足设定条件，则此序列被加入待输出序列数据集中。并记录序列中PTS含量数据和序列总长度到分析数据集中。循环处理直到文件结尾。如果待输出序列数据集不为空则将待输出序列数据集以fasta格式保存，将分析数据集以json格式保存到文本文件。

**使用：**将待筛选序列的fasta格式数据库拷贝到程序所在路径，执行本初步筛选程序。如果有符合条件的输出数据，该数据会以[result+原文件名]为文件名保存在程序所在路径。分析结果保存在程序所在路径下名为[compare+原文件名.txt]文件中。

* PartB:粘液素域猜测程序

取得程序所在路径下所有fasta格式数据库，遍历每数据库中序列。从每序列第一个氨基酸开始以8个连续氨基酸为基本单位将总蛋白逐步切片，游标+8。

**处理逻辑：**

**切片过程**：

* A如当前切片最后一氨基酸为PTS钟任意一个：
  + a下一氨基酸仍为氨基酸PTS中任意一个，则将总蛋白中下一氨基酸加入当前切片，游标+1。循环步骤A
  + b 下一氨基酸不是PTS中任意一个。保存当前切片到返回值中。进行下一个切片。
* B如当前切片最后一氨基酸不是PTS中任意一个：
  + c 将切片中最后一个氨基酸从切片中删除，游标-1.。循环步骤B直到当前切片长度为1或当前切片最后氨基酸是PtS中任意一个。

循环切片过程直到序列结束，将每切片保存在列表中返回调用者。

**组合目标域过程：**

**由于组合过程比较复杂，在此只做简介，不进行详细文字叙述。具体流程请参考源程序。**

**组合目标域的设计原则：以贪婪模式尝试组合每序列所有连续切片，连续的切片组合后如PTS含量满足设定条件且组合后切片长度满足长度下限则记录为可能的粘液素域。**

**组合处理步骤完成后将每序列中可能的域在输出结果中高亮显示。**

**使用：**将待筛选序列的fasta格式数据库(须经初步筛选后的)拷贝到程序所在路径，执行本程序。如果有符合条件的输出数据，该数据会以[原文件名.html]为文件名保存在程序所在路径。

源代码：<https://github.com/nopw/ProteinMucinDomainSearch>