

SeqCounter

关于

SeqCounter是一个DNA序列分析器。

本项目以[GPL3.0](#)协议开源，请务必遵守。

GitHub仓库地址：<https://github.com/nnrj/SeqCounter>

使用说明

运行

- 方式一 无参运行
 - 将所有待检测序列放入根目录下的 seqs 文件夹中；
 - 运行 SeqCounter.exe 或 SeqCounter.py，统计结果将保存于根目录下 results 文件夹中的 result_XXXXXXXX.log 文件中。
- 方式二 有参运行

参数	含义	作用	默认值	别名	备注
-i	输入文件之路径	指定待检测序列目录	./seqs/	-ifile	
-o	输出文件之路径	指定检测结果保存目录	./results/	-ofile	
-c	约束文件之路径	指定用于比对序列类型约束文件	./ini/virusinfo.ini	-cfile	
-- v	查看版本号	查看版本号		-- version	
-- h	查看帮助信息	查看帮助信息		--help	

约定

- 输入文件
 - 输入文件为序列文本，默认为 .txt 格式；
 - 所有输入文件应当放在默认或参数指定的输入文件路径下。

本程序根目录下的 tools 文件夹中提供了两个批处理工具：

fastaToTxt.bat：将 .fasta 格式的文件批量转换为 .txt；

txtToFasta.bat：将 .txt 格式的文件批量转换为 .fasta。

使用方法：

将需要转换的文件放入同一个文件夹，复制相应的 `xxx.bat` 文件到该文件夹，运行 `xxx.bat` 即可。

- 输出文件
 - 输出文件为普通文本或Excel文件，默认为 `.log` 格式的普通文本文件；
 - 输出文件中保存了检测结果；
 - 输出文件的名称为 `result_xxxxxxx.log`，其中 `xxxxxxx` 代表结果生成时的时间，精确到分钟。
- 序列类型约束文件
 - 序列类型约束文件（下文简称约束文件），为普通文本，以 `.ini` 为扩展名；
 - 默认为存储于根目录下的 `./ini/virusinfo.ini`；
 - 约束文件按特定格式，给出了序列类型与长度的对应，以供程序判断序列类型；
 - 具体格式规定如下：
 - 每个一行，病毒名与长度用英文连字符 `-` 连接；
 - 对于同一个病毒，多个长度，使用斜杠 `/` 隔开；
 - 除最后一行外，每行要以英文逗号结尾。
 - 合法约束格式如下（举例说明）：

```
1 合法约束格式如下：
2      病毒名A-长度，
3      病毒名B-长度1/长度2，
4      病毒名3-长度
```

配置文件说明

配置文件为程序根目录下的 `ini/setting.json`。

您可以通过修改配置文件来控制SeqCounter的功能开关及其他行为。

默认配置文件

```
1  {
2      "version": "2.1.5",
3      "seqCounter": {
4          "encoding": "utf-8",
5          "inputOptions": {
6              "seqPath": "./seqs/",
7              "seqExtensionName": ".txt",
8              "encoding": "utf-8",
9              "symbols": [">", ">>"]
10         },
11         "outputOptions": {
12             "resultPath": "./results/",
13             "resultExtensionName": ".log",
14             "encoding": "utf-8",
15             "compare": true,
16             "combineCompare": false,
17             "extractSeq": true,
18             "singleExtract": false,
19             "extractExtensionName": ".fasta",
20             "removeSymbols": [" ", "\n", "\t", "@num", " "],
```

```

21         "ignoreEmptySeq": true
22     },
23     "constraintOptions": {
24         "seqTypeList": "./ini/virusinfo.ini",
25         "seqTypeCheck": true,
26         "removeSymbols": [" ", "\n", "\t"]
27     }
28 }
29 }

```

含义说明

```

1  {
2      "version": "2.1.5", // 版本号
3      "seqCounter": { // 序列统计模块配置
4          "encoding": "utf-8", // 编码格式
5          "inputOptions": { // 输入文件配置
6              "seqPath": "./seqs/", // 输入文件目录（必须为目录）
7              "seqExtensionName": ".txt", // 输入文件拓展名
8              "encoding": "utf-8", // 输入文件编码
9              "symbols": [ ">", ">>" ] // 文件内序列分隔符
10         },
11         "outputOptions": { // 输出文件配置
12             "resultPath": "./results/", // 输出文件目录（必须为目录）
13             "resultExtensionName": ".log", // 输出文件拓展名
14             "encoding": "utf-8", // 输出文件编码
15             "compare": true, // 是否标识相同序列（true, 是; false: 否）
16             "combineCompare": false, // 是否跨文件对比（true, 是; false: 否）
17             "extractSeq": true, // 是否提取序列（true, 是; false: 否）
18             "singleExtract": false, // 是否单独提取（true, 是; false: 否）
19             "extractExtensionName": ".fasta", // 提取序列的拓展名
20             "removeSymbols": [" ", "\n", "\t", "@num", " "], // 提取序列时要移
                除的字符
21             "ignoreEmptySeq": true // 标识相同序列时，忽略空序列（true, 是;
                false: 否）
22         },
23         "constraintOptions": { // 约束文件配置
24             "seqTypeList": "./ini/virusinfo.ini", // 序列类型列表文件
25             "seqTypeCheck": true // 是否执行序列类型判断（true, 是; false: 否）
26         }
27     }
28 }

```

- removeSymbols: 提取序列时要移除的字符
 - 值为一个字符串数组;
 - 数组内容为要从序列中移除的所有字符或字符串;
 - 支持python的转义字符;
 - 支持批处理操作, 格式为 @批处理名称, 目前支持的批处理如下:
 - 移除所有数字: @num

若要移除的字符串恰好是批处理操作符, 可使用@符转义。

例如，要移除 @num，可在列表中指定 @@num。

升级日志

- 版本 2.1.5
 - 标识相同序列时，允许跳过空序列；
 - 修复类型判断选项配置不起作用的BUG。
- 版本 2.1.3
 - 序列提取时，支持自定义要去除的字符或字符串。
- 版本 2.1.2
 - 修复自定义输入文件名无法识别的BUG。
- 版本 2.1.1
 - 支持提取序列。
- 版本 2.1.0
 - 支持标识相同序列；
 - 修复配置文件开关无效的BUG；
 - 重构结果打印方法。
- 版本 2.0.2
 - 增加 --v（查看版本号）命令；
 - 增加 --h（查看帮助信息）命令。
- 版本 2.0.1
 - 修复 --c 命令不起作用的BUG。
- 版本 2.0.0
 - 支持参数控制；
 - 支持自定义输入输出文件；
 - 打包为 .exe 文件。
- 版本 1.8.0
 - 修复了病毒类型判断失效的BUG。

许可协议

见根目录下的[LICENSE](#)(可使用文本文档打开)。

开发团队

天河何处、木落

SeqCounter开发团队

2022年5月28日