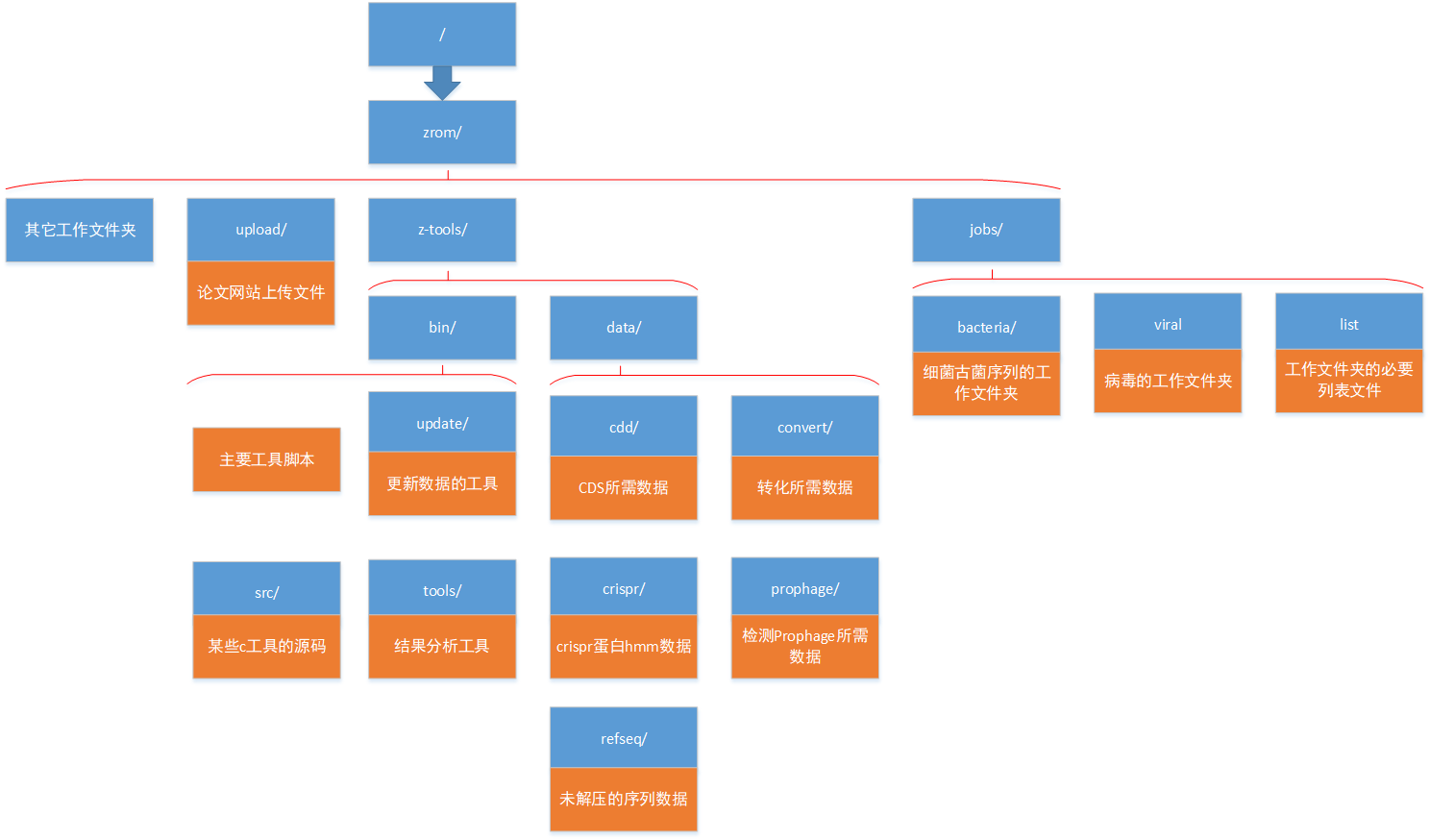
生信工具集Z-TOOLS使用文档

（赵世家工作部分）

作者：赵世家

时间：2017/8/31

1. 整体介绍
2. 工具集适用的工作
3. 通过北大服务器从NCBI获取目前已知细菌、古菌、病毒的序列。
4. 通过北大服务器从NCBI获取GI和REFSEQ的转换数据文件。
5. 通过北大服务器从NCBI获取CDSearch的数据文件。
6. 分析序列的spacer。
7. 分析序列的crispr区域。
8. 分析序列的selftarget情况。
9. 尝试翻译DNA序列成为蛋白质序列。
10. 分析序列获得crispr蛋白。
11. \*获得Prophage区域。
12. \*统计分析。
13. \*服务器负载限制。
14. 主服务器文件夹结构介绍



1. 必要环境介绍

Blast工具

Hmmer 3.0工具

Piler-cr工具

FragGeneScan工具

Linux Ubuntu 16.04下测试通过

注意/zrom/z-tools/bin要在环境变量中

1. 脚本详细介绍
2. 获取未解压的序列文件(一般不运行，除非官方数据库大批改变，或者本地未解压数据丢失，更新代价较大)

/z-tools/bin/update/

gain\_cdd\_data.py 获取cdsearch的数据文件（未解压），目标文件夹为/z-tools/data/cdd

update\_refseq\_\*.sh 获取对应的序列文件（未解压），目标文件夹为/z-tools /data/refseq/

1. 将未解压的序列文件解压到当前所在目录

/z-tools/bin/

get\_archaea 解压古菌序列，具体用法见脚本

get\_bact 解压细菌序列，具体用法见脚本

get\_viral 解压病毒序列，具体用法见脚本

PS：如果是分析所有序列的pipeline，需要海彬的分割脚本seperatefna，分割到zrom/jobs底下

1. 提取spacers以及Crispr区域的寻找

/z-tools/bin/

Findcrispr使用方法：

findcrispr 线程数 列表文件

列表文件每一列以制表符”\t”分割，第一列必须是REFSEQ，在脚本文件的fpath中定义寻找路径规则，分析文件名必须为\*.fna，输出文件在目标文件目录下，.spc为spacer序列文件，.csp为crispr区域

1. 预测crispr蛋白

/z-tools/bin/

get\_crisprsys.py 使用方法：

get\_crisprsys.py fna文件

输出.cpt文件为预测出的cas蛋白，.faa文件为蛋白质序列文件到目标文件目录下

get\_exists\_crisprsys.py 使用方法

get\_exists\_crisprsys.py 线程数量 列表文件

列表文件由上一个脚本（findcrispr）提供

这个脚本是用来对jobs底下的文件进行分析，也就是调用上面get\_crisprsys.py这个脚本，来对所有的序列进行分析

1. 预测Prophage区域

/z-tools/bin/

Prophagefinder 使用方法：

Prophagefinder faa文件

Faa文件由上一个脚本提供

1. 流程规划举例

每一个任务根据不同的需求在原有的脚本上需要作出更改

例子：

/zrom/simon/contig\_anal

希望分析一个contig组序列文件final.contigs.fasta ，找出里面所有的spacer序列，crispr区域，以及可能的cas蛋白。

1.首先运行./separate.py

分割final.contigs.fasta文件到contigs文件夹下，生成列表文件contig\_list

2.再运行./findcrispr 5 contig\_list

用5个线程分析contig\_list列表文件中的目标，得到spacer序列和Crispr区域信息，得到有Crispr区域的列表文件cripsrbacteria

（这里的findcrispr脚本里的fpath函数已经更改）

3.最后运行./exec\_anal cripsrbacteria，得到cas蛋白的预测结果

（exec\_anal脚本中调用get\_crisprsys脚本）