# GMBHS使用手册

目录

[GMBHS使用手册 1](#_Toc1653498330)

[一、 软件简介 1](#_Toc1746231235)

[二、 软件运行环境 2](#_Toc994868375)

[（一） 操作系统 2](#_Toc582368828)

[（二） Python 3](#_Toc459036109)

[（三） BLAST 4](#_Toc1354355738)

[三、 使用说明 5](#_Toc175349499)

[（一） 基本用法： 5](#_Toc1726265474)

[（二） 参数列表： 5](#_Toc95306991)

[（三） 参数详解 6](#_Toc419211199)

[（四） 使用案例 7](#_Toc1938181232)

[（五） 软件运行情况和结果解释 9](#_Toc2146492220)

## 软件简介

对于非模式生物的基因组草图序列来说，同源搜索是一项非常重要的注释信息来源，而且准确性很高。一般同源搜索获得的高相似位置都是可能同源的编码蛋白基因所在；然而它有一个非常重要的缺陷，就是无法获得精确的基因结构，包括上下游基因特征信号、以及高相似内部区域的外显子内含子剪接识别位点等。故而，我们在同源搜索的基础之上，基于Python编写一款自动批处理软件。该软件根据同源搜索的结果，提取相应的基因组序列，进而使用基因建模软件对其进行结构建模，以获取精确的基因结构模型；然后对建模输出的蛋白进行进一步鉴别；最后将结果整合并以gff3格式输出；该格式文件可直接用于基因组可视化工具，如：IGV或JBrowse等。

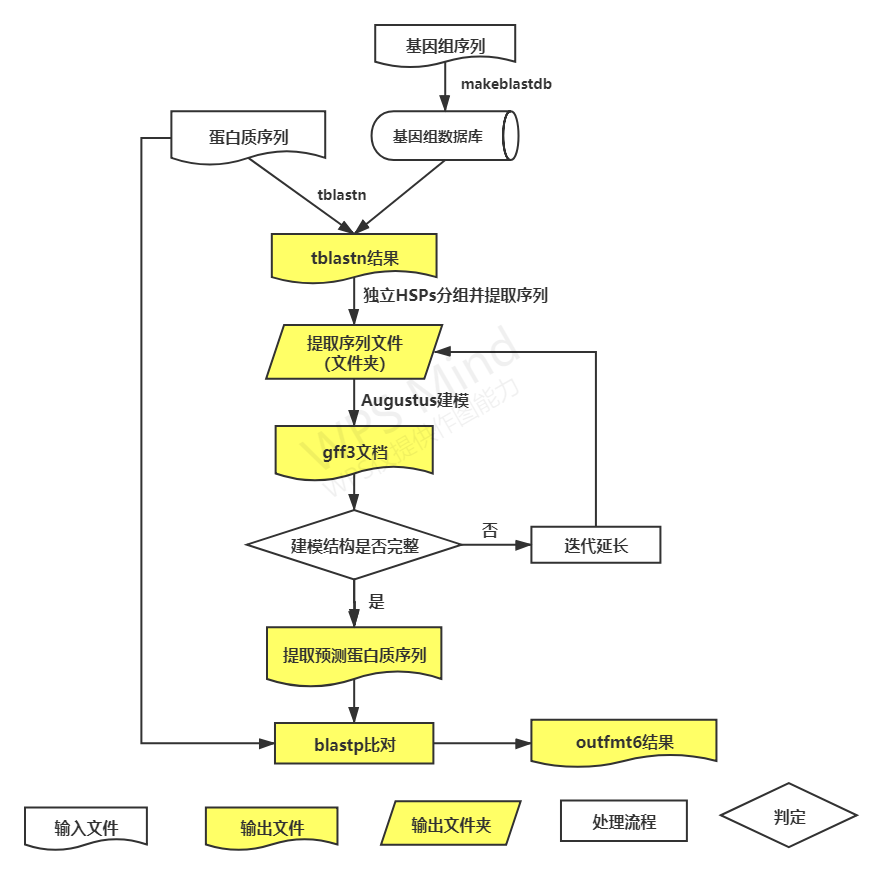


图1、软件处理流程图

## 软件运行环境

### 操作系统

支持Ubuntu16以及以上发行版本。

### Python

1. **版本：**

python3.0及以上版本。

1. **必需的Python扩展包和模块：**
2. sys：命令行参数
3. subprocess：系统命令
4. pathlib：路径、文件、目录
5. re：正则表达式
6. gzip：gz解压
7. tarfile：tar解压
8. zipfile：zip解压
9. unrar：rar解压
10. biopython（SeqIO）：解析和读取fasta序列
11. argparse：参数传入和解析
12. getopt：命令行参数

注：除biopython的SeqIO模块和unrar模块需要单独安装外，其他均为python3内置模块。unrar只有--genome\_remote下载的基因组文件为rar的压缩格式时会用到。

1. **安装：**
2. **python3的安装**

访问https://www.python.org/downloads/source/，选择适用于 Unix/Linux 的源码压缩包。下载及解压压缩包 Python-3.x.x.tgz，3.x.x 为你下载的对应版本号，以 Python3.8.5 版本为例。

#下载python3.8.5安装包

cd /opt

wget <https://www.python.org/ftp/python/3.8.5/Python-3.8.5.tgz>

#解压

tar -zxvf Python-3.8.5.tgz

#编译

cd Python-3.8.5

./configure

make && make install

#检查 Python3 是否正常可用

python3 -V

# Python 3.8.5

# 环境变量配制,/opt/python为python安装路径

# csh shell:

setenv PATH "$PATH:/opt/python"

# bash shell (Linux)：

export PATH="$PATH:/opt/python"

# sh 或者 ksh shell:

PATH="$PATH:/opt/python"

1. **Bio的SeqIO模块安装：**

使用pip安装即可，在终端输入命令pip install biopython。

1. **unrar的安装：**

使用pip安装即可，在终端输入命令：pip install unrar。

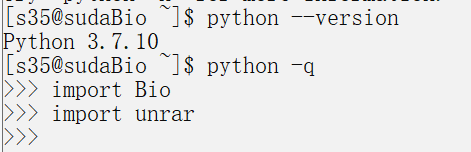


图2、python3及其关键模块安装成功的测试截图

### BLAST

**1、版本：**

ncbi-blast+2.5及以上版本。

1. **下载和安装：**

从其官网（<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>）下载blast最新版本。其中有不同系统不同类型的安装包，64位的linux下载ncbi-blast-2.11.0+-x64-linux.tar.gz。

# 安装包下载，以2.11.0版本为例

wget <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>ncbi-blast-2.11.0+-x64-linux.tar.gz

# 移动到要安装的目录,假如为/opt/

mv ncbi-blast-2.11.0+-x64-linux.tar.gz /opt/ncbi-blast-2.11.0+-x64-linux.tar.gz

cd /opt

#解压

tar -zxvf ncbi-blast-2.11.0+-x64-linux.tar.gz

# 环境变量配制

# 解压好的目录中bin目录的绝对路径:/opt/blast-2.11.0+/bin

echo "export PATH=/opt/blast-2.11.0+/bin:\$PATH" >> ~/.bashrc

source ~/.bashrc

# 查看是否安装成功

blastn -version

#如果显示版本信息，说明安装成功

注：对于部分Linux发行版可直接用包管理器安装，如 Debian和Ubuntu，直接使用命令sudo apt install ncbi-blast+即可安装成功。

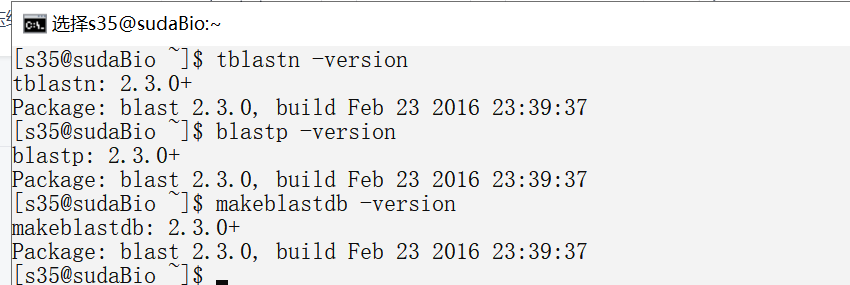


图3、ncbi-blast+安装成功的测试截图

## 使用说明

### **基本用法：**

GMBHS.py --protein proteins.fasta --genome genome.fasta [--step 50] [--outdir output]

注：[]代表该选项是可选的。

### 参数列表：

表1、GMBHS命令行参数列表

| **参数** | **说明** |
| --- | --- |
| -h,--help | 输出参数说明 |
| -p,--protein | 已知蛋白质序列文件(format:xxx.fasta)，可运行多个蛋白序列 |
| -g,--genome | 基因组序列文件 (format:xxx\_genome.fasta) |
| -w,--genome\_remote | 下载基因组序列的网址链接 |
| -s,--step | 迭代提取基因组序列的步长 |
| --score | 结果过滤中的相似性分值阈值，默认100 |
| --identity | 结果过滤中的一致性比例阈值，默认50 |
| --distance | 结果整合时的距离阈值，默认1000bp |
| --local\_modeling\_tool | 建模所使用的工具，默认为augustus |
| --max\_iteration | 延伸序列最大迭代次数，默认100 |
| --num\_threads | blast比对时设置的所用线程数 默认为1 |
| -o,--outdir | 结果输出目录,默认使用./output/作为输出目录 |

注：参数赋值尽量避免使用中文字符，否则可能导致程序出现异常错误。

用户在命令行中可以通过-h或--help参数获取所有参数的简要说明信息（图4）。

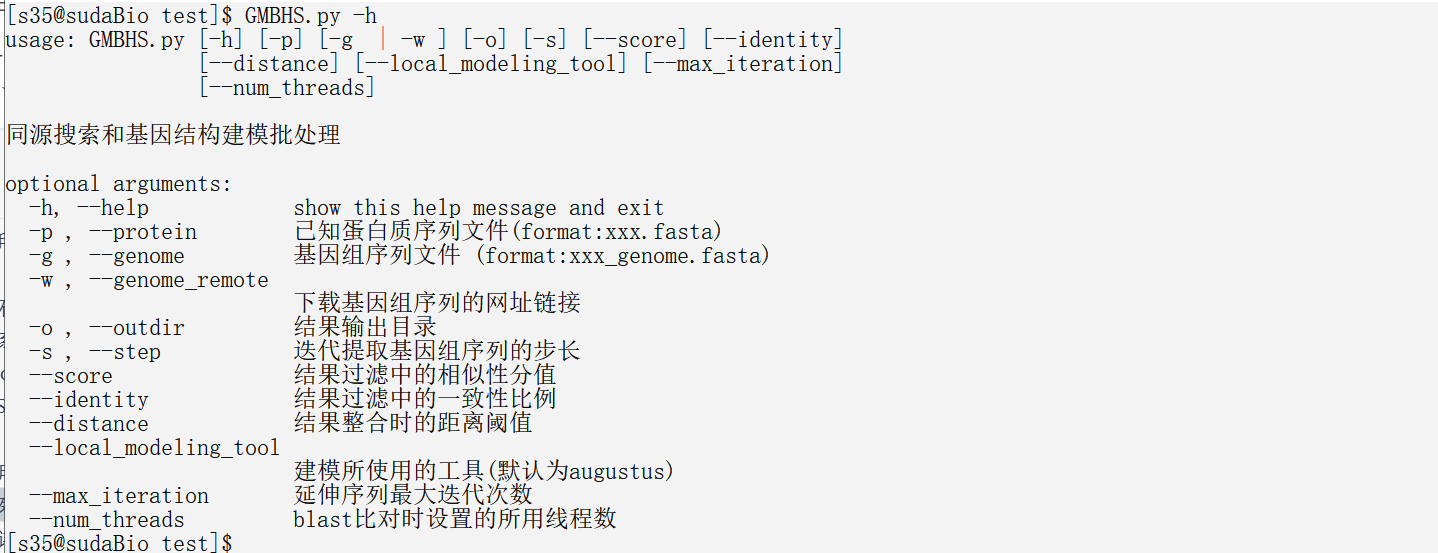


图4、命令行执行-h或--help参数的屏幕回显截图

### 参数详解

1. **已知蛋白质序列（必选）**

**-p, --protein：**设定已知蛋白质序列文件，文件后缀名为fasta格式，该文件可包含单个序列，也可包含多条蛋白质序列。

1. **基因组序列文件（必选）**

**-g, --genome：**设定基因组序列文件；一个单一的存放基因组序列信息的fasta格式文件。程序接收基因组序列文件后调用makeblastdb程序创建参考基因组数据库。

**-w, --genome\_remote：**设定基因组序列文件远程下载地址；用户可利用该参数提供一个参考基因组序列文件下载地址，程序可以自动下载、解压，并调用makeblastdb程序创建参考基因组数据库。

**以上两个参数任选其一，若设定了多个参数，则程序会弹出错误提示。**

1. **迭代步长（可选）**

**-s, --step:**对tblastn比对得到的基因组区间，上下游各延伸指定长度，然后提取相应的基因组序列，用于后续基因结构建模。

1. **相似打分阈值（可选）**

**--score：**设定tblastn结果筛选的相似打分阈值，该值大于0，默认值100；低于该值的结果将被删除。

1. **一致性阈值（可选）**

**--identity：**设定tblastn筛选的一致性阈值，取值范围0-100，默认值50；低于该值的结果将被删除。

1. **结果整合时的距离阈值（可选）**

**--distance：**对高相似的HSPs区间进行整合，默认值1000bp。对于同一个蛋白的多个HSPs，如果间距（--distance）低于阈值，则认为其可能一个基因的不同外显子，故而将其跨越区间合并，作为该基因的跨越区间。

1. **建模所使用的工具（可选）**

**--local\_modeling\_tool：**选定基因序列预测建模的工具，目前仅支持augustus建模工具。默认为augustus。

1. **延伸序列最大迭代次数（可选）**

**--max\_iteration：**为防止一直建模不成功情况，也不能一直循环迭代下去，设置最大迭代次数（默认为100），若超出了最大迭代次数，则弹出警告并终止程序。

1. **blast比对时设置的所用线程数（可选）**

**--num\_threads：**设置tblastn比对时所使用的线程数。默认值为1，当用户输入的线程数大于blast所限定的最大线程数，则使用blast限定的最大线程数进行比对。

1. **输出目录（可选）**

**-o, --outdir：**设定程序计算结果的输出目录，默认为当前工作路径下的output目录；如果设定的目录不存在，程序会自动创建；如果已经存在，则直接输出到该目录。

### 使用案例

1. **案例一：**

准备好保存单个蛋白的序列文件（α-1，3葡聚糖合成酶蛋白序列.fasta）和基因组序列文件在当前工作目录下，设置迭代提取基因组序列的步长为50，输出目录为GMBHS\_output1，其他参数使用默认值进行测试。首先对基因组序列进行建库，将建库结果保存到输出目录中，之后进行tblastn将蛋白序列和基因组进行比对，输出比对结果(\_tblastn.outfmt6)。根据比对结果对HSPs进行分组和整合，并根据提示区间从基因组中提取序列并同时使用augustus进行建模，判断结构完整性和迭代提取序列。之后对预测建模好的完整基因结构编码蛋白进行提取，提取后的预测蛋白和原蛋白进行blastp比对，若比对结果符合要求即保存augustus预测结果和预测出的蛋白序列。

GMBHS.py \

--protein ./α-1，3葡聚糖合成酶蛋白序列.fasta \

--genome CCTCCM2012259\_genome.fasta \

--output /home/student/s35/1830401138/yan/GMBHS\_output1 \

--step 50 --score 100 --identity 50 --distance 1000 \

--local\_modeling\_tool augustus --max\_iteration 100 --num\_threads 1

1. **案例二：**

在当前工作目录下存在原始蛋白质序列文档（包含多个蛋白质序列），但无基因组序列文件和本地基因组库，用户可在命令行参数中输入基因组文件下载地址，即可在联网的情况下远程下载基因组序列并进行建库，之后进行同案例一的同源蛋白比对和建模以及鉴定工作。

GMBHS.py \

--protein ./all\_proteins.fasta \

--genome\_remote <https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/336/255/GCA_003336255.1_ASM333625v1/GCA_003336255.1_ASM333625v1_genomic.fna.gz>

--output /home/student/s35/1830401138/yan/GMBHS\_output2 \

--step 50 --score 100 --identity 50 --distance 1000 \

--local\_modeling\_tool augustus --max\_iteration 100 --num\_threads 10

1. **案例三：**

在当前工作目录下存在原始蛋白质序列文档和基因组序列文件，无输出文档参数。默认在当前工作目录下创建输出目录。设置迭代提取序列步长为100，相似性分值阈值设置为150，一致性比例阈值为60%，建模工具设定为Augustus，blast线程数设置为5，其他参数保持默认参数进行测试。

GMBHS.py \

--protein ./all\_proteins.fasta \

--genome CCTCCM2012259\_genome.fasta \

--step 100 \

--score 150 \

--identity 60 \

--distance 1000 \

--local\_modeling\_tool augustus \

--max\_iteration 100 \

--num\_threads 5 \

### 软件运行情况和结果解释

**（1）软件运行情况：**

以上面描述的案例1为例介绍软件运行情况和输出的结果。当在命令行终端，运行本软件进行转录本数据处理时，软件会将每一个关键处理过程的关键信息通过终端窗口回显向用户反馈（图5），以便用户实时查看软件处理进程、以及可能的出错或警告信息，从而及时作出调整、修正等

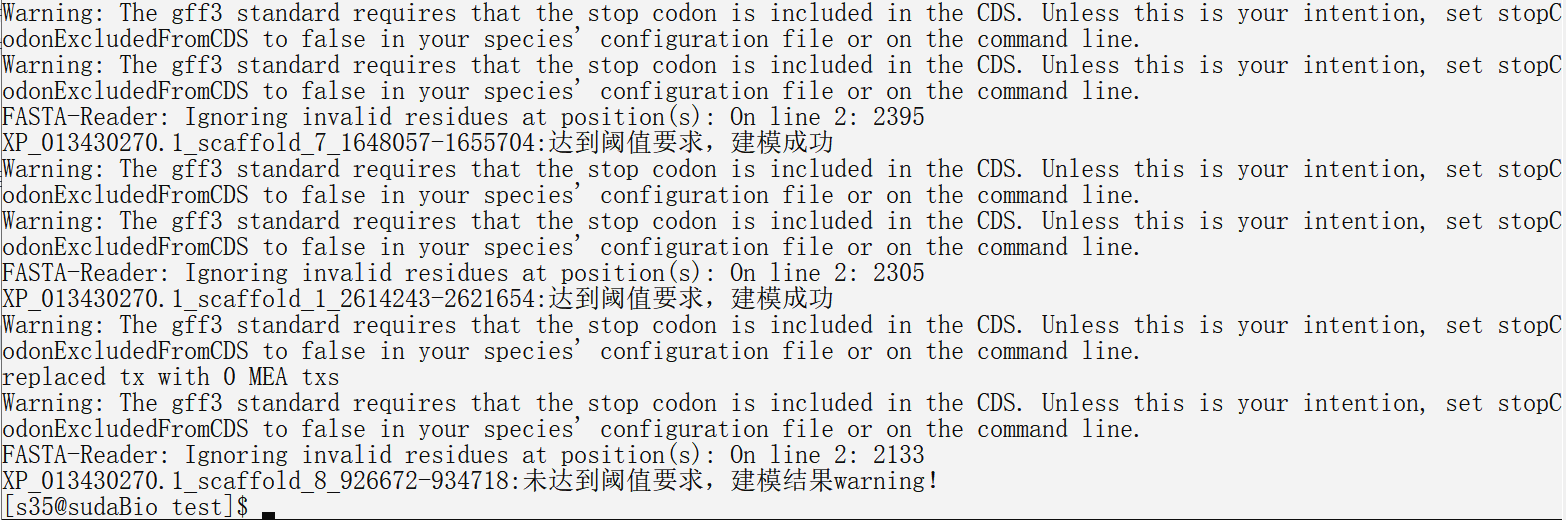
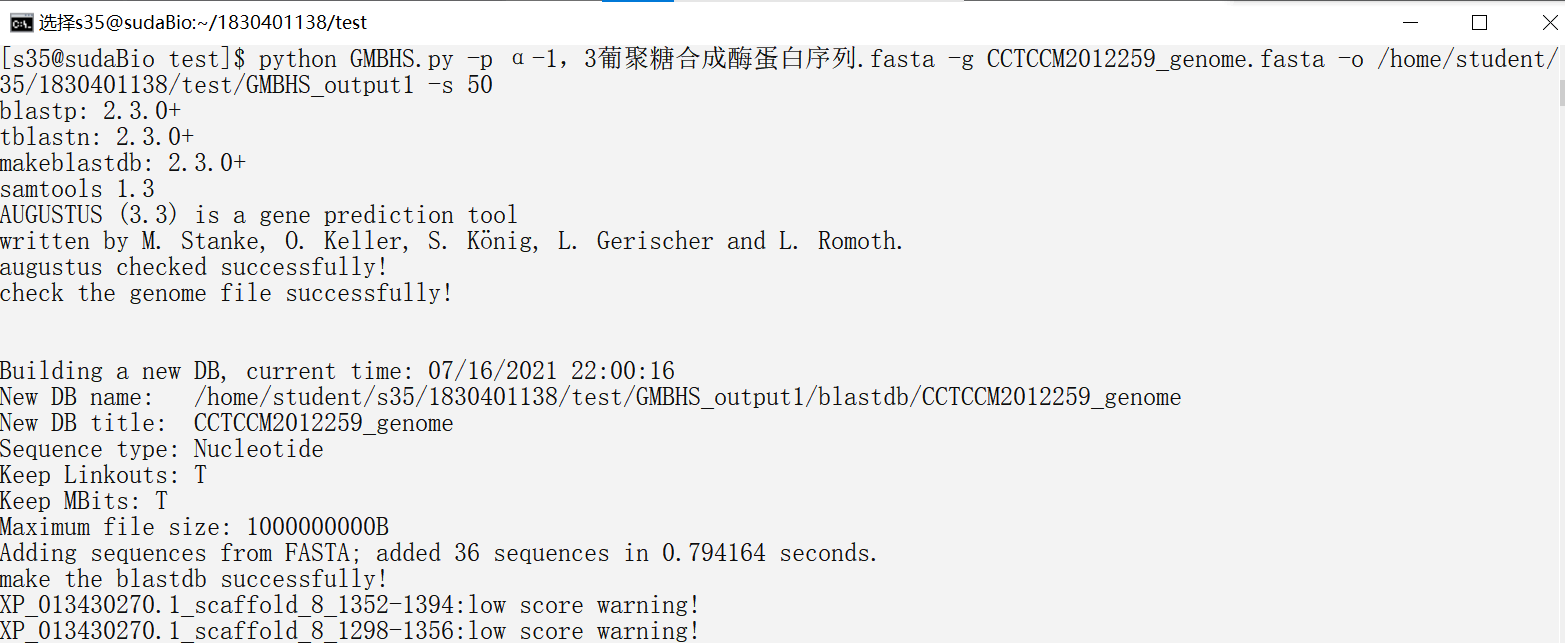
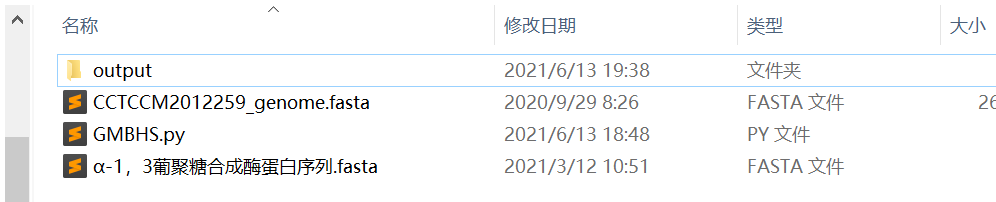


图5、案例1执行的命令行回显截图

**（2）软件运行结果：**

程序执行时，会在当前工作路径下创建结果输出目录test，并在处理过程中，在该目录下保存上述计算过程中输出的各种中间和最终结果文件（图6）。该结果输出目录的结构、以及其中的结果文件描述详见表2。



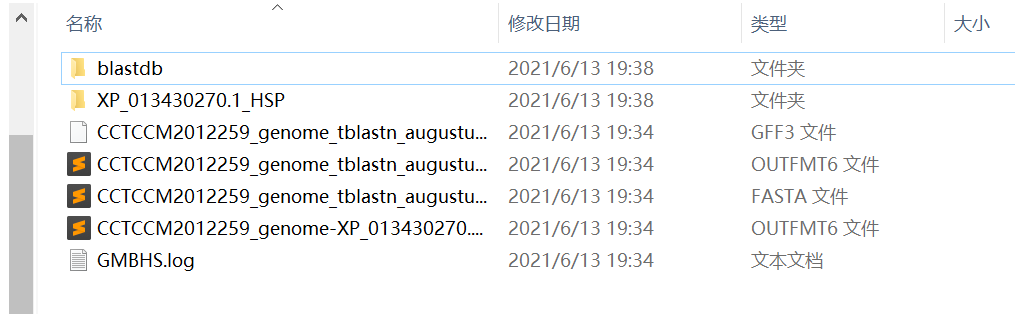






图6、案例1执行结果输出目录和结果文件

表2、软件运行结果输出目录和文档的结构和描述

|  |  |
| --- | --- |
| **结果文档和目录结构** | **描述** |
| --GMBHS.log | 程序运行的日志文件 |
| --blastdb  |--CCTCCM2012259\_genome.nsq  |--CCTCCM2012259\_genome.nin  |--CCTCCM2012259\_genome.nhr | 基因组建库结果 |
| --（蛋白质ID）\_HSP  |--scaffold\_8\_926722-934644.fasta  |--scaffold\_8\_926672-934694.fasta  |--scaffold\_7\_1648188-1655654.fasta  |--scaffold\_7\_1648138-1655704.fasta  |--scaffold\_1\_2614293-2621562.fasta | 蛋白1在基因组中匹配并分组好的HSPs序列文件。 |
| --CCTCCM2012259\_genome-（蛋白ID）\_tblastn.outfmt6 | 原始蛋白和基因组比tblastn比对结果（outfmt6格式） |
| --CCTCCM2012259\_genome\_tblastn\_augustus.gff3 | 根据tblstn比对到HSP区间提取序列后使用augustus建模的gff3格式注释结果 |
| --CCTCCM2012259\_genome\_tblastn\_augustus\_pred.fasta | 从gff3建模结果中提取的预测蛋白序列文件。 |
| --CCTCCM2012259\_genome\_tblastn\_augustus\_blastp.outfmt6 | 预测蛋白序列和原始蛋白blastp比对结果 |