《生物信息技能训练》实验记录

学　院: 医学部

专　业: 生物信息学

姓　名: 朱泽峰

学　号: 1730416009

题　目: 全基因组基因的从头预测及结构建模

组　号: 小组 2

组　长: 朱泽峰

组　员: 李定洋、裘或然、张书凡、郑宇翔

2020 年 9 月 15 日

全基因组基因的从头预测及结构建模

摘要

DNA 序列分析是后基因组时代计算生物学的一个重要领域。从上世纪九十年代至今，单单从基因组序列中进行基因结构从头算预测的计算方法在很大程度上促进了研究者对各种生物学问题的理解。虽然这方面的计算预测已经有了不少方法与对应软件，但如何为研究对象 (数据) 选择合适的方法、数据集下开发的软件，从而做到较为稳健与准确地预测也面临着挑战。

本次实验就挖掘多种模式生物基因组中的序列特征及其在基因预测中的应用进行比较性实验。主要目的是对不同的主流预测方法/软件在不同种的模式生物基因组数据集下的预测结果进行校验，并试图阐明区别所在。

关键词：全基因组从头预测; 全基因组结构建模; 模式生物；非模式生物

目录

1 实验设计 **4**
1.1 前期调研 4
1.1.1 Organism 4
1.1.2 Software 5
1.2 实验流程设计 5
1.2.1 任务分工 6

2 软硬件 **7**

3 实验步骤 **7**
3.1 GenBank 数据库选择与下载选定物种的基因组序列和注释文档 7
3.2 UniprotKB 数据库下载选定物种所在分类的所有已知蛋白 8
3.3 基因预测软件的检索、下载与安装 8
3.3.1 Augustus 的下载与安装 (by 郑宇翔) 8
3.3.2 GeneMark-ES 的下载与安装 (by 裘 yu 然) 9
3.3.3 Geneid 的下载与安装 (by 郑宇翔) 10
3.4 全基因组序列的基因从头预测与结构建模 (by 郑宇翔、裘 yu 然) 11
3.4.1 Augustus 的基因从头预测与结构建模 (by 裘 yu 然、酵母 by 郑宇翔) . 11
3.4.2 GeneMark-ES 的基因从头预测与结构建模（by 裘 yu 然） 12
3.4.3 Geneid 的基因从头预测与结构建模（by 郑宇翔） 13
3.5 预测结果的比较、挑选与整合（部分完成 by 李定洋、朱泽峰） 14
3.5.1 各软件 GFF 文件经过重注释的结果与原 GFF 文件进行对比 14
3.6 预测结果的序列提取以及比对分析与改良 (by 张书凡、朱泽峰) 15
3.7 本地 BLAST 数据库的创建与同源基因搜索做格式转换（by 李定洋） 15
3.8 用对应程序将 blast 结果转化为 gff3 格式（by 李定洋） 15
3.9 JBrowse 的和 IGV 的注释信息可视化（by 张书凡） 15

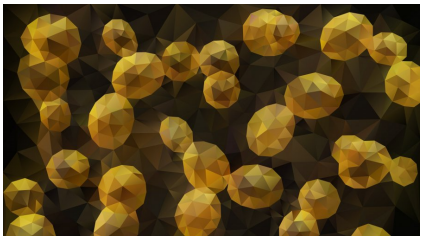
1 实验设计

1.1 前期调研

1.1.1 Organism

鉴于此次实验目的在于比较不同软件对不同物种尤其的模式生物以及非模式生物的预测结果的差异，我们组经过调研，决定选用一种常被用于数据训练的模式生物，一种略少见的模式生物以及一种非模式生物。在考虑了数据可用性以及实验可行性方面，最终确定如下物种:

1. 酿酒酵母 (Saccharomyces cerevisiae) (真菌)
2. 秀丽线虫 (Caenorhabditis elegans) (线虫)
3. 草履虫 (Paramecium tetraurelia) (纤毛虫)



(a) Saccharomyces cerevisiae

(b) Caenorhabditis elegans

(c) Paramecium tetraurelia

图 1: 所选物种

选定理由如下: 酿酒酵母 (Saccharomyces cerevisiae) 是最简单的真核生物之一，其基因组长度为 12,157,105(1 千万级) 个碱基对，包含 6692 个基因 (Ensembl); 秀丽隐杆线虫 (Caenorhabditis elegans) 的基因组长度为 1 亿个碱基对，包含的基因数量与人类相似，约为 20500 个基因 (Ensembl)；草履虫 (Paramecium tetraurelia) 具有两种细胞核: 微核和大核，大约 87Mbp(Ensembl)。

同时，我们小组成员在实验正式开始之前已于数据库初步检索有无对应的数据资源，具体调研分工如下:

1. GeneBank: 裘或然, 张书凡
2. Ensembl: 朱泽峰
3. UniProt: 郑宇翔
4. 调研软件可用度/有无失效: 李定洋, 张书凡
5. 文献调研 [2]: 朱泽峰

1.1.2 Software

根据相关文献 [3] 的记录与比较, 我们起初选择了 Augustus、GeneMark-ES 与 GeneScan 进行基因从头预测。但在实际使用过程中，我们发现 GeneScan 的本地化相对难以使用。因此我们最终决定将 GeneScan 更换为 Geneid。最终所选软件如下:

1. Augustus
2. GeneMark-ES
3. Geneid

1.2 实验流程设计

1. 于 GenBank 数据库选择与下载下列物种的基因组序列和注释文档
   (a) Saccharomyces cerevisiae
   (b) Caenorhabditis elegans
   (c) Paramecium tetraurelia
2. 于 UniprotKB 数据库下载这些物种所在分类的所有已知蛋白 (排除该物种自身的已知蛋白)
3. 基因预测软件：安排两人对 Augustus 等软件检索相关的评测比较的文章，并选取若干个工具，安装使用测试，解决软件使用过程中出现的问题
4. 使用上一步选定的多个基因预测软件，对全基因组序列进行基因预测和结构建模，结果转成 GFF3 格式
5. 利用 gffcompare 软件比较不同软件的预测结果，进行挑选与整合
   (a) 解读 gffcompare 结果

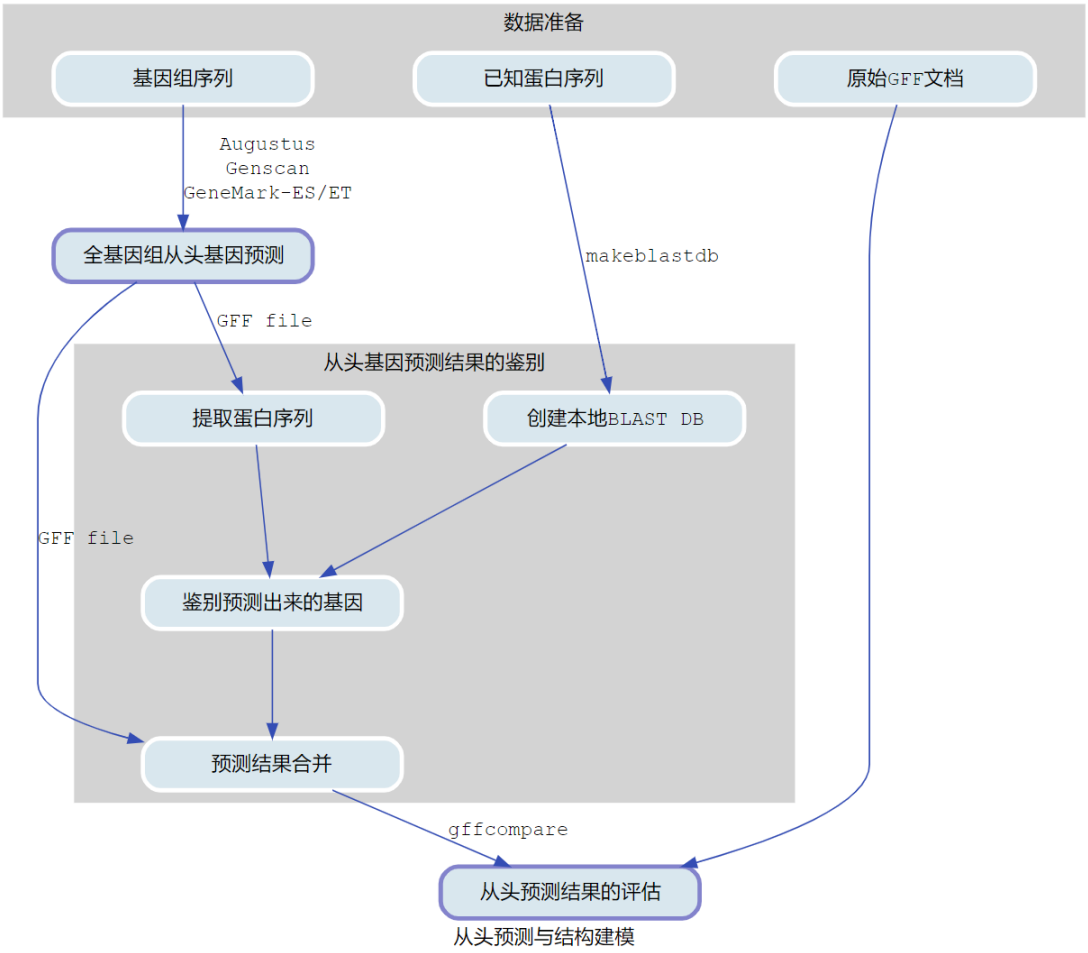


图 2: 实验流程图

(b) 两两共同特征的选择: e.g. 局部重合, 完全重合

(c) 而后进行挑选与整合

6. 将 4 和 5 的结果与 1 下载的原始注释文档进行比对，分析异同，评估优劣，改良

7. 本地 blast 配置，用 blast 比对鉴别第 4 和 5 步的结果中输出的蛋白质

8. 用 blast92gff3.pl 程序转化 blast 结果为 gff3 格式 (修改参数)

9. 利用 JBrowse 和 IGV 将注释组的注释信息可视化，比较两者优劣

1.2.1 任务分工

1,2 : 张书凡

3,4 : 裘或然, 郑宇翔

5, : 李定洋, 朱泽峰

6, : 朱泽峰, 张书凡

7,8 : 李定洋

9, : 张书凡

2 软硬件

操作系统与运行环境等参数

1. 操作平台：阿里云服务器
2. 平台配置：双核 2.40GHz CPU,4GiB 内存
3. 操作系统：Ubuntu 20.04 (64bit)
4. 使用软件：
   (a) Augustus 3.3
   (b) GeneMark-ES ver 4.*
   (c) Geneid 1.2
   (d) Gffread
5. 编程语言
   (a) Python3
   (b) Perl
   (c) R
   (d) Shell Script

3 实验步骤

3.1 GenBank 数据库选择与下载选定物种的基因组序列和注释文档

从 NCBI 的 Genome 数据库下载草履虫、秀丽线虫、酵母菌的基因组序列和 GFF 格式的注释信息，搜索地址如下：

1. 酵母菌（Saccharomyces cerevisiae）

(a) https://www.ncbi.nlm.nih.gov/genome/?term=Saccharomyces+cerevisiae

2. 秀丽线虫（Caenorhabditis elegans）

(a) https://www.ncbi.nlm.nih.gov/genome/?term=Caenorhabditis+elegans

3. 草履虫（Paramecium tetraurelia）

(a) https://www.ncbi.nlm.nih.gov/genome/275

3.2 UniprotKB 数据库下载选定物种所在分类的所有已知蛋白

从 UniPort 数据库下载下面三个物种所在分类的所有已知蛋白（排除该物种自身的已知蛋白质）。草履虫属于纤毛亚门，秀丽线虫属于线虫门，酵母菌属于真菌界。搜索关键词如下：

1. 酵母：‘taxonomy:fungi NOT "saccharomyces cerevisiae" AND reviewed:yes‘

2. 线虫：‘taxonomy:nematoda NOT "caenorhabditis elegans" AND reviewed:yes

3. 草履虫：‘taxonomy:ciliophora NOT "paramecium tetraurelia" AND reviewed:yes‘

3.3 基因预测软件的检索、下载与安装

选定如下软件的理由参见上述。

在其他小组成员分别利用 Augustus,Geneid,GeneMark 三个物种的基因组数据进行分析后，得到了软件输出的 GFF 与 GTF 格式文件。预先设定的分析流程是按照 Augustus 的 GFF 输出文件内容来安排，即提取其中的蛋白序列信息于建立好的 blast 数据库进行搜索，找到对应的 UniProt 条目进而重注释例如 Augustus 输出的结果，方便后续进行对比分析。但是在实际实验过程中发现，Geneid,GeneMark 最新版本输出结果中 (GFF3 格式或 GTF 格式文件) 没有所需的蛋白序列信息。因此安排负责同学对软件参数与版本进行调查，最终决定回退至特定版本，以获得序列信息。(by 朱泽峰)

3.3.1 Augustus 的下载与安装 (by 郑宇翔)

软件下载:(http://bioinf.uni-greifswald.de/augustus/binaries)

在 Github 或者 Augustus 的官网下载地址中获得最新版本的 Augustus 软件 [4]，由于使用的是云服务器，可有多种方式下载安装 Augustus，但 wget 或从 git 上 clone 的方式受限于服务器带宽，于是选择在官网下载地址直接下载至本地，并上传至服务器。

进入服务器的软件压缩包存放目录执行以下代码：

```
# 解压缩
$ tar zxf augustus-3.3.3.tar.gz
$ cd augustus-3.3.3
$ cd src
# 编译安装
$ make
```

安装完毕后可以将 Augustus 路径添加至环境变量中 (包括 Config 路径)，方便使用：

```
export AUGUSTUS_CONFIG_PATH=/root/augustus-3.3.3/config/
export PATH=$PATH:/root/augustus-3.3.3/bin:/root/augustus-3.3.3/scripts
```

直接输入 augustus 测试可用性：

```
root@iZ2ze8zvy13uv2pb8yq4gcZ:~# augustus
AUGUSTUS (3.3.3) is a gene prediction tool
written by M. Stanke, O. Keller, S. König, L. Gerischer and L. Romoth.

usage:
augustus [parameters] --species=SPECIES queryfilename

'queryfilename' is the filename (including relative path) to the file containing
    the query sequence(s)
in fasta format.

SPECIES is an identifier for the species. Use --species=help to see a list.
# 参数下略
```

3.3.2 GeneMark-ES 的下载与安装 (by 裘 yu 然)

软件下载:(http://topaz.gatech.edu/GeneMark/license_download.cgi)

在 GeneMark 的官网上找到下载地址，选择对应的版本和操作系统并填写网页下方的表单信息（对学术用户免费），得到软件压缩包和用户 key[3]。

将安装包与 key 一同上传于服务器中，解压缩安装包后执行以下命令导入 key(同时确保 key 文件在用户根目录下)：

```
$ cp gm_key ~/.gm_key
```

由于 GeneMark 使用了 Perl 语言编写，需要安装相关的 Perl 模块，根据其安装说明文

档则有如下模块需要安装:

```
The following Perl modules are required:
   YAML
   Hash::Merge
   Logger::Simple
   Parallel::ForkManager
   MCE::Mutex
   Thread::Queue
   threads
```

同时也提供了多种安装方式的建议，这里我们使用了 Cpanm 模块来进行模块的安装，安装示例如下:

```
$ cpan App::cpanminus
$ cpanm YAML
```

在所有依赖安装完毕且 key 设置成功的情况下，运行软件目录下的 check_install.bash 来确认安装是否有模块缺失和错误:

```
root@iZ2ze0yrfbj6dp6da7ebwvZ:~/gmes_linux_64# sh check_install.bash
Checking GeneMark-ES installation
All required components for GeneMark-ES were found
```

根据信息显示所有依赖模块都已安装完毕，可进行下一步测试。

测试可用性（by 裘或然）

在软件目录存在一个用于测试的 GeneMark-E-tests 目录，用于测试包括 ES 步骤在内的 EP 功能是否可用，根据其内部的说明文档来运行这个例子:

```
# 创建测试目录,并执行GeneMark-EP(使用核心数根据CPU实际情况进行调整)
mkdir test; cd test
../../../gmes_petap.pl --seq ../input/genome.fasta –EP --dbep
    ../input/proteins.fasta --verbose --cores=2 --max\_intergenic 10000
# 如果软件安装正常，则会在output文件夹中输出genemark.gtf结果文件
```

3.3.3 Geneid 的下载与安装 (by 郑宇翔)

软件下载:(https://github.com/guigolab/geneid)

该软件的安装步骤较为简略，直接在 Geneid 的 git 页面中即可找到安装说明，将 Geneid 的压缩包 clone 至服务器或直接下载后上传，按照以下步骤完成安装 [1]:

```
tar -zxvf geneid.tar.gz
# 移动至geneid根目录
# 编译geneid
make
# 测试并查看geneid的参数帮助
bin/geneid -h
NAME
       geneid - a program to annotate genomic sequences
SYNOPSIS
       geneid [-bdaefitxsz]
              [-D] [-Z]
              [-G] [-X] [-M] [-m]
              [-WCF] [-o]
              [-O <gff_exons_file>]
              [-R <gff_annotation-file>]
              [-S <gff_homology_file>]
              [-P <parameter_file>]
              [-E exonweight]
              [-Bv] [-h]
              <locus_seq_in_fasta_format>
RELEASE
       geneid v 1.2a
# 剩余参数显示略
```

3.4 全基因组序列的基因从头预测与结构建模 (by 郑宇翔、裘 yu 然)

3.4.1 Augustus 的基因从头预测与结构建模 (by 裘 yu 然、酵母 by 郑宇翔)

将三个物种的原始基因组序列文件上传至服务器，分别放在独立的文件夹中准备进行基因从头预测步骤，执行基因预测的代码如下：

```
nohup augustus --gff3=on --outfile=Sc_augustus_out.gff3
    --species=saccharomyces_cerevisiae_S288C --stopCodonExcludedFromCDS=FALSE
    Sac_cerevisiae.fna &

nohup augustus --gff3=on --outfile=elegant_augustus_out.gff3
    --species=caenorhabditis --stopCodonExcludedFromCDS=FALSE cae_elegans.fna &
```

```
nohup augustus --gff3=on --outfile=para_augustus_out.gff3 --species=tetrahymena
    --stopCodonExcludedFromCDS=FALSE paramecium_tetra.fna &
```

问题解决（by 裘或然）：

在尝试运行之前，遇到了草履虫在备选的物种列表中不存在的问题。但是在备选列表中找到了和草履虫亲缘关系很近的四膜虫（tetrahymena），并且四膜虫与草履虫等其他纤毛虫一样，具有双元核型（nuclear dimorphism），因此选择了这个物种参数进行替代。

代码运行过程中，我们还遇到了 Segmentation Fault 的报错问题，于是我们在 git 的 issue 中寻找，发现很多用户也出现了类似的情况。起初认为是内存溢出问题，但在选用了两条染色体执行的情况下依旧会出现这个报错。最后经过排查是系统版本影响，在 Ubuntu 18.04(64bit) 版本的情况下基因组规模较大的情况就会出现该错误，于是最后换用了 Ubuntu 20.04(64bit) 版本成功重新执行代码。

结果整理（by 裘或然）：

由于参数设置了输出格式为 gff3，并且文件中带有蛋白质序列信息（重要），可以直接等待后一步骤的提取，将结果分别存放至独立的文件夹中等待下一步操作。Augustus 的最终预测结果基因数目如下：

表 1: Augustus 最终预测结果基因数目

物种	酵母菌	秀丽线虫	草履虫
Augustus 预测基因数目	5465	6445	12983

3.4.2 GeneMark-ES 的基因从头预测与结构建模（by 裘 yu 然）

文件准备与基因预测执行（by 裘或然）：

将 3 个物种的基因组序列文件置于独立文件夹中，运行 GeneMark-ES 来执行基因预测：

```
# 真菌需要添加独立的—fungus参数
nohup ../gmes_petap.pl --seq paramecium_tetra.fna --ES --verbose --cores=2
    --max_intergenic 10000 &
nohup ../gmes_petap.pl --seq cae_elegans.fna --ES --verbose --cores=2
    --max_intergenic 10000 &
nohup ../gmes_petap.pl --seq Sac_cerevisiae.fna --ES --fungus --verbose --cores=2
    --max_intergenic 10000 &
```

结果整理（by 裘或然）：

在文件夹中会生成一系列结果，其中生成需要的结果文件为 genemark.gtf，GeneMark-ES 生成 gtf 文件相较于 gff3 文件有一定程度的差异，且并不含有蛋白质序列的具体信息，

只有位置信息。因此需要用到 GeneMark-ES 文件夹中的 get_sequence_from_GTF.pl 来结合原始基因组序列来根据序列段获得蛋白质序列，并输出蛋白质序列结果文件。但是在使用该代码的时其一部分有关正则表达式的代码有不适用输出 gtf 文件的情况，经过修改后可以正常使用。

对于三个物种的基因预测量分别如下：

表 2: GeneMark-ES 最终预测结果基因数目

物种	酵母菌	秀丽线虫	草履虫
GMES 预测基因数目	5471	22591	13800

最后经过处理可从 gtf 与 fasta 文件中提取出 prot_seq.faa 蛋白质序列文件和 nuc_seq.fna。另外 gtf 在 gffcompare 中无法直接对比，因此需要进行转换。

3.4.3 Geneid 的基因从头预测与结构建模（by 郑宇翔）

该软件需要训练好的物种配置文件，geneid 的网站上已经有了部分物种经训练获得的部分物种的配置文件。选择的配置文件如下所示：

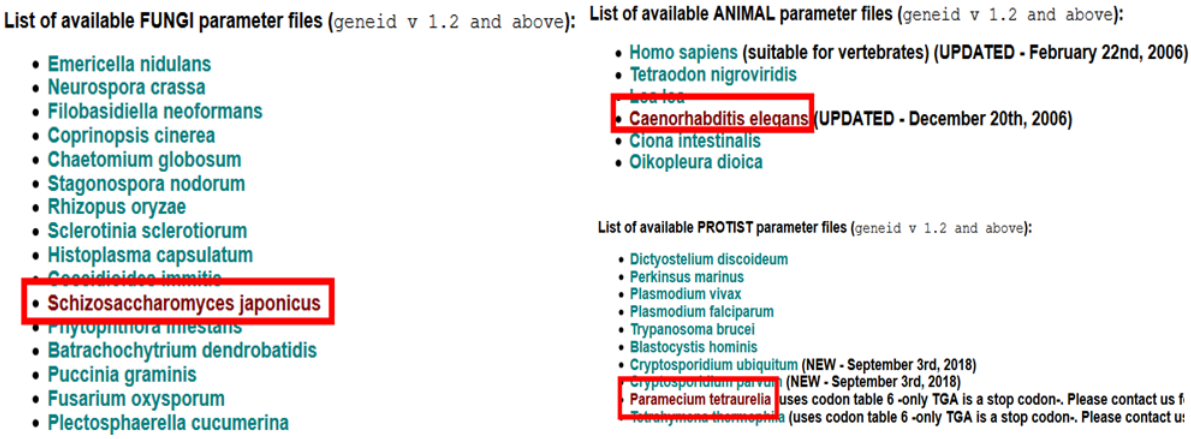


图 3: geneid 配置文件

其中，酿酒酵母 (Saccharomyces cerevisiae) 使用的是日本裂殖酵母 (Schizosaccharomyces japonicus) 的配置文件。而草履虫的配置文件只考虑 TGA 作为终止密码子。这可能会对最终预测的基因数量和准确度产生影响。

使用 geneid 1.4 分别对三个物种进行预测，输入如下命令：

```
../geneid -3P sjaponicus.param_Oct_12_2006 Sac_cerevisiae.fna > Sac_cerevisiae.gff3
../geneid -3P ptetraurelia.param.Mar_5_2005 paramecium_tetra.fna >
   paramecium_tetra.gff3
```

```
../geneid -3P celegans.param.Dec_20_2006 cae_elegans.fna > cae_elegans.gff3
```

从而得到 gff3 格式文件。

但是该 gff3 文件没有包含蛋白质序列，尝试寻找可以输出蛋白质序列的方法。经查找资料，发现在 1.4 版本中，geneid 去掉了可以直接输出蛋白质序列的功能，只有在较早版本（如 1.2）中，才有可以直接输出序列的功能，且只能在软件自带的 geneid 格式文件中输出。使用 geneid 1.2 分别对三个物种进行预测，输入如下命令：

```
../geneid -vP sjaponicus.param_Oct_12_2006 Sac_cerevisiae.fna >
   Sac_cerevisiae.geneid
../geneid -vP ptetraurelia.param.Mar_5_2005 paramecium_tetra.fna >
   paramecium_tetra.geneid
../geneid -vP celegans.param.Dec_20_2006 cae_elegans.fna > cae_elegans.geneid
```

另行编写 python 脚本将蛋白质序列从 geneid 文件中提取出来。预测得到的基因数量如下所示：

表 3: Geneid 最终预测结果基因数目

物种	酵母菌	秀丽线虫	草履虫
Geneid 预测基因数目	4073	5705	23386

根据目前的研究结果，酿酒酵母共有 6275 个基因, 其中可能约有 5800 个真正具有功能；秀丽隐杆线虫有大约 20000 个基因；而草履虫有大约 4 万个基因。酿酒酵母和秀丽隐杆线虫的结果均较为接近，但 geneid 只预测出了 4073 个草履虫的基因，结果远少于实际情况，这可能是因为提供的模型只考虑 TGA 作为终止密码子所致。

3.5 预测结果的比较、挑选与整合（部分完成 by 李定洋、朱泽峰）

3.5.1 各软件 GFF 文件经过重注释的结果与原 GFF 文件进行对比

首先是安装最新版 gffcompare 软件:

```
$ git clone https://github.com/gpertea/gffcompare
$ cd gffcompare
$ make release
```

接着分别设定各个物种的参考 GFF 文件以及各个软件的对应 GFF 结果文件，运行 gffcompare, 命令运行示例如下:

```
$ gffcompare -V -r $gff_sc_r $gff_sc_au -o augustus_compare_sc
```

```
Prefix for output files: augustus_compare_sc
Loading reference transcripts..
  6445 reference transcripts loaded.
  1 duplicate reference transcripts discarded.
Warning: adjusted transcript g1570.t1 boundaries according to terminal exons.
Warning: adjusted transcript g2180.t1 boundaries according to terminal exons.
  5465 query transfrags loaded.
Cleaning up..
Done.
```

输出结果示例如下:

```
$ ls augustus_compare_sc*
augustus_compare_sc.annotated.gtf augustus_compare_sc.stats
augustus_compare_sc.loci        augustus_compare_sc.tracking
```

结果分析: ...
可以看到:

- 对于 Saccharomyces cerevisiae，除去内含子层面的预测，Augustus 在 Precision 层面 (即给出的预测结果的正确率) 以及 Sensitivity 层面 (即正确预测出的结果于参考 GFF 中的占比) 的预测都足够理想
- 对于 Caenorhabditis elegans，在 Sensitivity 层面的预测不够良好，即正确预测出的结果于参考 GFF 中的占比较低，且转录本层面的 Precision 也不够理想
- 对于 Paramecium tetraurelia，仅 BaseLevel 的 Precision 数值较高，其余数值都很低，说明 Augustus 选用的模型对于本次实验的数据材料存在比较明显的错配

3.6 预测结果的序列提取以及比对分析与改良 (by 张书凡、朱泽峰)

3.7 本地 BLAST 数据库的创建与同源基因搜索做格式转换（by 李定洋）

3.8 用对应程序将 blast 结果转化为 gff3 格式（by 李定洋）

3.9 JBrowse 的和 IGV 的注释信息可视化（by 张书凡）

参考文献

[1] Enrique Blanco, Genís Parra, and Roderic Guigó. Using geneid to identify genes. *Current Protocols in Bioinformatics*, 18(1):4.3.1–4.3.28, 2007.

[2] Deng F. Huang Y, Chen SY. Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. *Comput Struct Biotechnol J.*, (14):298–303, 2016 Jul 27.

[3] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20):6494–6506, 01 2005.

[4] Mario Stanke and Burkhard Morgenstern. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, 33(suppl$_2$) : $W465 - -W467, 072005$.