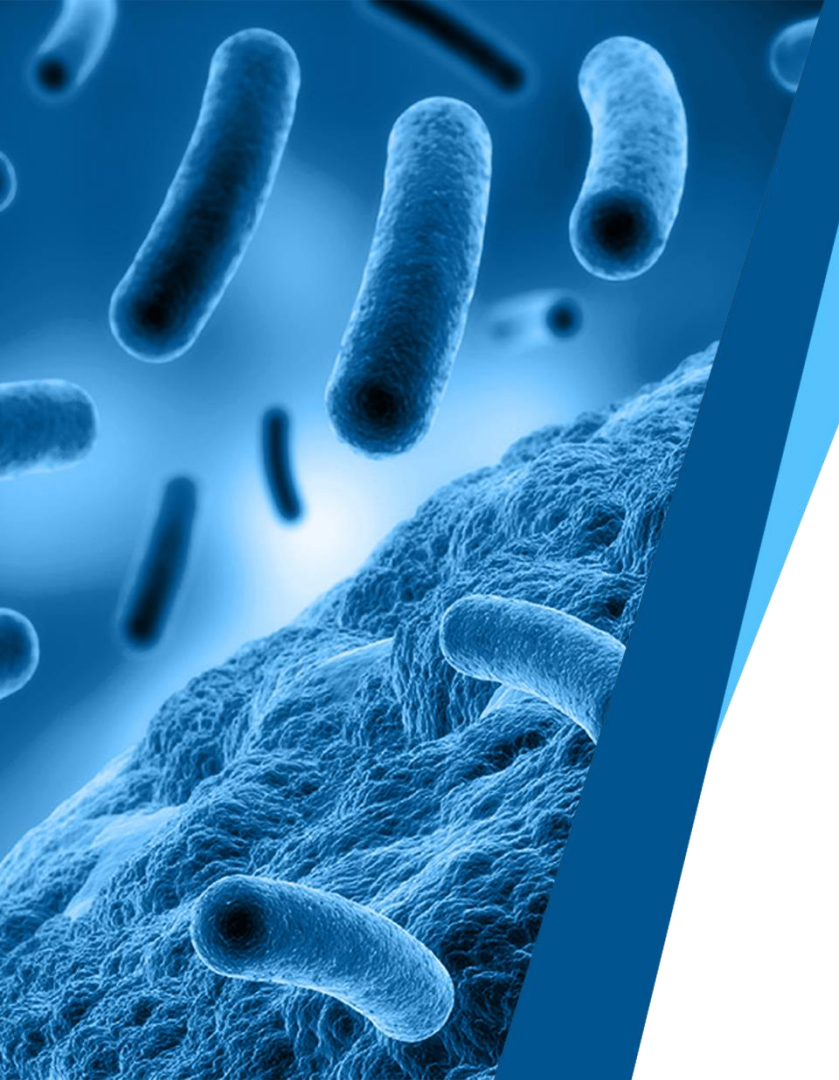

国家微生物科学数据中心
National Microbiology Data Center

Prodigal介绍及使用

Prodigal简介

国家微生物科学数据中心
National Microbiology Data Center

- Prodigal是为细菌和古菌基因组进行蛋白编码基因预测的软件,其缩写源于PROkaryotic DYnamic Programming Genefinding ALgorithm,表示原核生物基因预测的动态规划算法
- 2010年,发表在BMC Bioinformatics
- 截至2021年2月，该出版物已被引用4400多次

原核生物基因特点

原核生物mRNA的特征：

- 1.原核生物mRNA的5’端无帽子结构，3’端没有或只有较短的多聚（A）结构
- 2.原核生物起始密码子AUG上游有一被称为Ribosome Binding Site (RBS)或SD序列（Shine –Dalgarno sequence）的保守区，该序列与16S-rRNA 3’端反向互补，被认为在核糖体-mRNA的结合过程中起作用

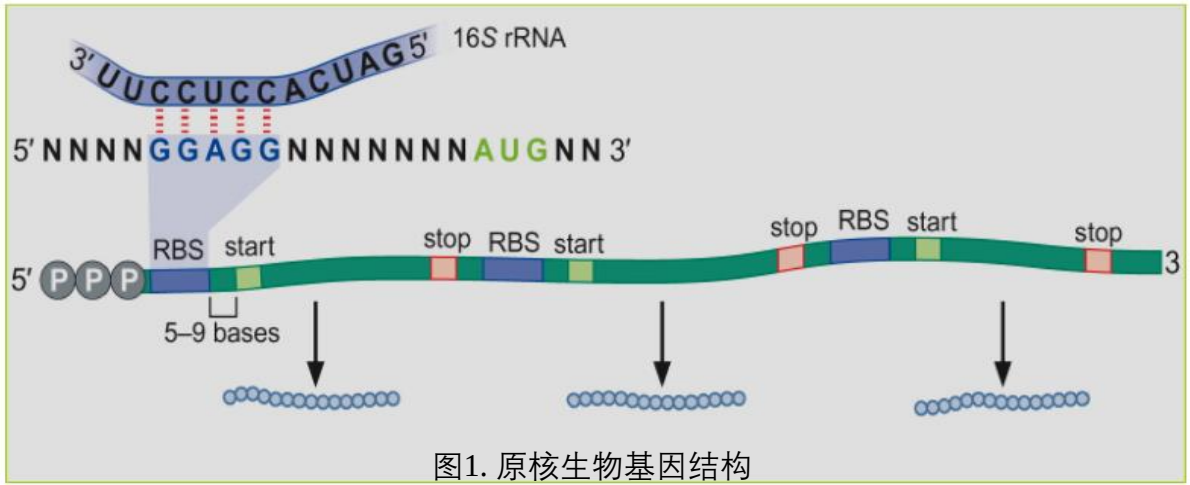

图1. 原核生物基因结构

原核生物基因特点

- 3.原核生物常以AUG（有时GUG，甚至UUG）作为起始密码子;真核生物几乎永远以AUG作为起始密码子
- 4.半衰期短：原核生物中，mRNA的转录和翻译是在同一个细胞空间里同步进行的，蛋白质合成往往在mRNA刚开始转录时就被引发了；大多数细菌mRNA在转录开始1分钟后就开始降解
- 5.许多原核生物mRNA以多顺反子的形式存在：原核细胞的mRNA(包括病毒)有时可以同时编码几个多肽
  - 单顺反子mRNA (monocistronic mRNA):只编码一个蛋白质的mRNA
  - 多顺反子mRNA(polycistronic mRNA):编码多个蛋白质的mRNA

Prodigal特点

Prodigal可以做的：

- 处理基因组类型多样: Finished genomes、draft genomes以及metagenomes
- 输出格式多样: GFF3、Genbank以及Sequin表格形式
- 运行速度快: 在PC上10秒钟就能分析出E. coli K-12基因组的蛋白序列
- 无监督学习: 使用无监督的机器学习算法，无需提供任何训练数据，而是从输入序列本身自动学习基因组的特性，包括遗传密码使用以及核糖体结合位点的motif识别
- 可以处理基因组gaps、contig、scaffolds和不完整基因预测

Prodigal特点

Prodigal不能做的：

- 预测RNA基因: Prodigal暂不能预测RNA基因
- 处理含有内含子的基因: 原核生物种含有内含子的基因非常罕见，所以Prodiga省去了对这一部分基因的预测
- 基因功能注释: Prodigal并不为它预测的基因提供功能注释
- 处理移码突变(frame shifts): 不包含任何处理碱基插入或删除的逻辑, Indel类型的测序错误将对基因预测产生影响
- 病毒基因预测: 慎重，尽管可以进行预测, 但是没有大规模验证

Prodigal安装

国家微生物科学数据中心
National Microbiology Data Center

1. Mac OS X安装

```
# 更新homebrew
brew update
# 克隆prodigal库
brew tap hyattpd/prodigal
# 安装
brew install prodigal
```

2. Linux 安装

```
conda install -c bioconda prodigal
```

3. Windows 安装

点击安装包按照提示安装

下载地址：https://github.com/hyattpd/Prodigal

Prodigal输入文件

多种类型的基因组文件：

- Finished genomes: 原核生物基因组完整基因组
- Draft genomes: 原核生物组装草图基因组
- Metagenomes：元基因组组装基因组

Prodigal适用于上述所有基因组的基因预测，并且支持多种格式的输入文件

Prodigal输入文件

国家微生物科学数据中心
National Microbiology Data Center

- 单个或多个序列的FASTA格式序列
- 单个或多个序列的Genbank格式序列
- 单个或多个序列的EMBL格式序列
- 推荐FASTA格式序列

序列ID　　序列注释

>gi|13650073|gb|AF349571.1 Homo sapiens hemoglobin alpha-1 globin chain (HBA1) mRNA, complete cds
CCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGACGACAAGACCAACGT
CAAGGCCGCCTGG
GGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGT
TCCTGTCCTTCCCCA　　序列
CCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGG
CCACGGCAAGAA
GGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGC
TGTCCGCCCTGAGC

图2. FASTA格式序列示例

Prodigal基本命令

常规基因预测：

```
prodigal -i my.genome.fna -o gene.coords.gbk -a protein.translations.faa
```

-i 输入文件，必须指定

-o 输出文件，默认gbk格式，若不写则标准输出到屏幕

-a 输出蛋白序列文件

Prodigal基本命令

更改输出文件格式：

prodigal -i my.genome.fna –f gff -o gene.gff3

-f, --output_format: 指定输出格式，可选如下：

gbk: Genbank-like format (Default)

gff: GFF format

sqn: Sequin feature table format

sco: Simple coordinate output

Prodigal基本命令

同时输出基因gff3、核酸以及蛋白：

```
prodigal -i my.genome.fna –f gff -o gene.gff3 –d  gene.fna –a gene.faa
```

-d 指定输出基因核酸序列文件

-a 指定输出基因蛋白序列文件

Prodigal基本命令

单个基因组和元基因组预测模式：

```
prodigal -i my.genome.fna –f gff -o gene.gff3 –p single
```

-p: 预测模式

single

meta

默认single

Prodigal基本命令

改变使用的密码子表：

prodigal -i my.genome.fna –f gff -o gene.gff3 –g 11

-g: 指定使用密码子表，数值1-25，默认11

11: Standard Bacteria/Archaea，标准细菌/古菌密码子表

4: Mycoplasma/Spiroplasma ，支原体和螺旋体密码子表

1-25：其他密码子表

Prodigal输出结果

国家微生物科学数据中心
National Microbiology Data Center

1. 基因坐标信息：GFF3

```
##gff-version  3
# Sequence Data: seqnum=1;seqlen=4674;seqhdr="contig_204037"
# Model Data: version=Prodigal.v2.6.3;run_type=Metagenomic;model="39|Rickettsia_conorii_Malish_7|B|32.4|11|1";
gc_cont=32.40;transl_table=11;uses_sd=1
contig_204037   Prodigal_v2.6.3 CDS     2       190     8.5     +       0       ID=1_1;partial=10;start_type=E
dge;rbs_motif=None;rbs_spacer=None;gc_cont=0.323;conf=87.70;score=8.54;cscore=5.32;sscore=3.22;rscore=0.00;usc
ore=0.00;tscore=3.22;
contig_204037   Prodigal_v2.6.3 CDS     490     1956    75.5    +       0       ID=1_2;partial=00;start_type=A
TG;rbs_motif=None;rbs_spacer=None;gc_cont=0.365;conf=100.00;score=74.84;cscore=74.06;sscore=0.79;rscore=-0.99;
uscore=-0.47;tscore=2.90;
contig_204037   Prodigal_v2.6.3 CDS     2112    2534    58.4    -       0       ID=1_3;partial=00;start_type=A
TG;rbs_motif=None;rbs_spacer=None;gc_cont=0.336;conf=100.00;score=59.14;cscore=53.38;sscore=5.75;rscore=-0.99;
uscore=3.07;tscore=2.90;
```

1.GFF3 header;

2.基因位置信息

图3. GFF3格式示例

Prodigal输出结果

国家微生物科学数据中心
National Microbiology Data Center

1. 基因坐标信息：GFF3

Order	Comment
1	**ID：每个基因的唯一标识符，由序列的序号和序列中该基因的序号组成（用下划线分隔）。例如，“ 4_1023”表示文件中第4个序列中的第1023个基因**
2	**partial：指示基因是否完整。“ 0”表示该基因具有完整边界，而“ 1”表示该基因在该边缘处“未完成”（即部分基因）。例如，“ 01”表示基因在右边界是部分基因，“ 11”表示两个边缘都不完整，“ 00”表示具有起始密码子和终止密码子的完整基因**
3	**start_type：起始密码子的序列（通常是ATG，GTG或TTG）。如果基因没有起始密码子，则该字段将标记为“Edge”**
4	**stop_type：终止密码子的序列（通常是TAA，TGA或TAG）。如果基因没有终止密码子，则该字段将标记为“Edge”**
5	rbs_motif：Prodigal发现的RBS motif（例如“ AGGA”或“ GGA”等）
6	rbs_spacer：起始密码子到RBS motif之间的碱基数
7	**gc_cont：基因序列的GC含量**
8	gc_skew：基因序列的GC Skew
9	**conf：基因的置信度得分，表示该基因为真基因的概率，即78.3%表示Prodigal认为该基因78.3%的概率真实基因，而21.7%的可能性为假阳性结果**
10	得分：该基因的总得分
11	cscore：六聚体编码部分得分，即该基因是真正蛋白质的得分
12	sscore：该基因翻译起始位点的分数；它是以下三个字段的总和：
13	rscore：该基因的RBS motif的分数
14	uscore：围绕起始密码子的序列的分数
15	tscore：起始密码子类型的分数（ATG vs. GTG vs. TTG vs. Nonstandard）
16	mscore：其余信号的得分（终止密码子类型和前导/滞后链信息）

Prodigal输出结果

国家微生物科学数据中心
National Microbiology Data Center

2. 基因核酸序列：fna

```
>contig_204037_1 # 2 # 190 # 1 # ID=1_1;partial=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.32
3
GACAATCAAGTAATACCTGAGATATCACCACTTTTCAGTGCGGGTATTTATTATACTGGAGAAGAATCTG
CATCTATCGGTATGGAAACAACTTTCACCTTATTCATGATTACGGTGGATTTTGCTGTCATTACAGTAAT
AGTCTATAGAAAAAAGAGAAAAAAGTTGAAAATAAAAAAAGCTGAATAA
>contig_204037_2 # 490 # 1956 # 1 # ID=1_2;partial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.
365
ATGATAATAATTTCTTATAGAATAAAAGCCAGTATTTATCAAACTCAATTAACAGATTTCATTGATATGA
AAAGTATTGAAAAGTACGACATGAATGTTCCAAAAGATTACAGGGTTGCTCGAAAAAAATTTCATTGGAT
CCCTATTCATCCTAATAGAGGGAAGCGAGCAAATCAAACTATGAGTGTAATGTCCAAATCTCAGAGAAAG
ATACATGATTTCATAGATTGGGTGCCTAAATTAAAGGATAAATACTTATACAGGTTTCCACGCGTAAAAA
TCCCTCGTTTCGTAGACTTTGGATTGGGAGTGATCTCTAAAATGGAATGGGATATGCATCGCTATCTTAA
AGGATTTAACGGGGTGTTTGAAAACTTATTTGAATTTAATGATTTTAGTTTTTTCCAAGAGTTACAAGGA
AACCTTGAGAATTGCGGTGTGTCATTTAAGGATATGTTTATCGAAGACTTACTTGCATACGAGATGTTGC
GGATTAACTTGGGCTTCAAAAACTATACTGGAATAGAAAGAATGGGTAAATTTCTCCTTTATCCTCCCTT
ATTCAACATTACTCATGACCCTAATTTCTTTCCAACAGCACAAGATATCAGTTACGTAATGACCAGAATC
CCCGCTGAAGCGTTATTTGAGTTTTTTCAGCTGTTAGTGAAAGAATGCATTGATTGTGGGATTATAGTCC
CAAGAATCCTCATCTGGGATGGACAATTCATACGCTCCAATTGCAGCAATAATAAAAAGAAAGGGAACAC
TAAATACAATGATCCCGATGCAGGTTATTGTAGACATAATGGTGTGAAAAAGGGTGTGGGCTATGATCCC
GGAATACTATATGTTCATTGTTTTAATCGCTGGTTACCTATCTATTTTAAGATGTTTGCTGGAAATCGGA
ACGATATTCTTGCGTTTAGAGAAACTATGGAAGCATTTTTTGCCAACACCGAGTACGAATGGCAAGTGGT
CATCGCAGATTCGGGACCATATTCTCTGCAAAATATGGAGAATATTCGGTCTAAAGGATTAATACCTATT
```

图4. 基因核酸序列

Prodigal输出结果

国家微生物科学数据中心
National Microbiology Data Center

3. 基因蛋白序列：faa

```
>contig_204037_1 # 2 # 190 # 1 # ID=1_1;partial=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.32
3
DNQVIPEISPLFSAGIYYTGEESASIGMETTFTLFMITVDFAVITVIVYRKKRKKLKIKK
AE*
>contig_204037_2 # 490 # 1956 # 1 # ID=1_2;partial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.
365
MIIISYRIKASIYQTQLTDFIDMKSIEKYDMNVPKDYRVARKKFHWIPIHPNRGKRANQT
MSVMSKSQRKIHDFIDWVPKLKDKYLYRFPRVKIPRFVDFGLGVISKMEWDMHRYLKGFN
GVFENLFEFNDFSFFQELQGNLENCGVSFKDMFIEDLLAYEMLRINLGFKNYTGIERMGK
FLLYPPLFNITHDPNFFPTAQDISYVMTRIPAEALFEFFQLLVKECIDCGIIVPRILIWD
GQFIRSNCSNNKKKGNTKYNDPDAGYCRHNGVKKGVGYDPGILYVHCFNRWLPIYFKMFA
GNRNDILAFRETMEAFFANTEYEWQVVIADSGPYSLQNMENIRSKGLIPIIRARKNLKTH
PVREFKKNFLFNTDYVPKEWSDEYLLKIYSFRPMIEQGNSYNNTFYNASRMNNRGMDAAI
KLRSIIYILELLKALTAYKLGRSDLIMKPTAFESSWYVNFRLALPRLAIQSGYKILSHNP
VLSRRNLL*
>contig_204037_3 # 2112 # 2534 # -1 # ID=1_3;partial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=
0.336
MIFFEELRVETPEREILLDITNELRKVVNKFDVKEGVCRIYIPHTTAGITINENADPSVK
KDISKFLNKLIPKGGGLGYSFKHGEGNSDAHIKCSLTGHSVEILIHDNRFMLGTWQGIMF
AEYDGPRNRRVYVQVQGESV*
```

图5. 基因蛋白序列

NMDC云平台演示

国家微生物科学数据中心
National Microbiology Data Center

国家微生物科学数据中心（NMDC）网址：

http://nmdc.cn

NMDC云平台演示

国家微生物科学数据中心
National Microbiology Data Center

1. 上传数据：文件系统-添加文件
2. 指定输入文件、工作目录
3. 设置运行参数
4. 运行
5. 在工作目录查看结果