

## Introduction

原核生物蛋白编码基因预测软件：提供快速准确地预测蛋白编码基因，格式包括GFF3, Genbank或Squin table format；可以处理组装好的基因组，或未组装好的基因组和宏基因组；用户可以指定存在gap的gene和如何处理contig边缘的genes；可以识别大部分gene的转录起点，且输出格式可包含基因组上所有潜在的转录起始点，以及对应置信值，RBS motif等。

Shine-Dalgarno sequence

[http://parts.igem.org/Help:Ribosome\\_Binding\\_Sites/Mechanism](http://parts.igem.org/Help:Ribosome_Binding_Sites/Mechanism)

细菌核糖体结合到mRNA特殊的序列位置，主要是核糖体结合位点和起始密码子。RBS和起始密码子之间的距离很重要，两者需要距离6-7个核酸，这样才能保证都能接触到核糖体复合体的合适位置。16s rRNA能够自由结合到包含序列"5'-ACCUGC-3'"的mRNA，该完整序列为Shine-Dalgrarno(SD)序列，能够在细菌mRNA上完整或者部分发现，为细菌和古生菌的信使RNA结合位点，一般启动密码AUG上游约8bp。该序列帮助招募核糖体到mRNA上，通过连接核糖体到起始密码子，开始蛋白合成。一旦完成招募，tRNA可在序列上根据密码子增加氨基酸，从转录起点开始向下游移动。

Genetic Code	11	4	1
Start codon (F-metionin)	ATG GTG TTG	ATG GTG TTG	ATG
Stop codon (translation termination)	TAA TAG TGA	TAA TAG	TAA TAG TGA

## Usage

指定预测模型，版本3.X

1. normal mode, 根据提供的序列，直接预测
2. anonymous mode, 根据提前计算训练的文件来预测输入序列
3. tarining mode, 同normal mode, 同时保留training文件供以后使用

-p, --mode: 默认single, 当个基因组, 可包含任何数目序列; train, 仅用于训练, 可以是一个或多个相近的基因组的mutltiple fasta文件; anon, 使用现有的训练文件预测, 针对metagenomic数据或单个短序列。

旧版PRODIGAL v2.6.3 [February, 2016], single对应normal, meta对应anon。

```
prodigal -i 38588-scaffolds.fasta -o 38588.coords.gbk -a 38588.protein.translations.faa
```

prokka使用记录: `prodigal -i prokka_analysis\43\fna -c -m -g 11 -p single -f sco -q`

-i, 指定输入文件, 可以为单个或多个fasta文件, genebank或embl格式, 推荐fasta格式

-o, 指定输出文件(gene coordinates)

-a, 指定输出转录的蛋白序列

-d, 指定输出对应核酸序列文件

-s, 指定完整的起始文件

-w, 输出统计文件

-f, 指定输出格式, gbk默认, gff, sqn(sequin feature table format), sco(simple coordinate output)

-g, 转录密码表, 11为细菌/古生菌; 4为支原体/螺原体, 默认先11后4

旧版: PRODIGAL v2.6.3 [February, 2016]

```
prodigal -i 38588-scaffolds.fasta -f gff -o 38588.coords.gff -a  
38588.protein.translations.faa -d 38588.nucelotides.fna -s 38588.stats
```

anonymous mode, 新版为 -p anon, 旧版 -p meta

```
prodigal -i 38588-scaffolds.fasta -p meta -a  
/Data_analysis/Ref_database/NCBI/Klebsiella_pneumoniae_ncbi/GCF_000240185.1_ASM240  
18v2_genomic.faa -o 38588.anon.gbk -p meta
```

training mode, 可以指定输出training文件供以后使用, 旧版无此功能

```
prodigal -i genome1.fna -p train -t genome1.trn
```

```
prodigal -i genome2.fna -t genome1.trn -o genome2.gbk -a genome2.faa
```

## 输出

ID: 序列经过排序的gene identifier, 4\_1023, 表示第4个序列的第1023个gene

partial: 表示gene是否超过了序列边缘, 0表示具有完整边界start和end, 1表示部分gene, 例如: 01表示gene部分存在右侧边界; 11表示gene两边超过边界; 00表示gene具有完整起始密码子

start\_type: ATG/GTG/TTG, 如果没有, 为Edge

stop\_type: TAA/TGA/TAG, 如果没有, 为Edge

rbs\_motif: RBS motif, 核糖体基序, 例如AGGA/GGA等

rbs\_spacer: motif和起始密码之间的距离

gc\_skew: gene序列的gc skew

conf: 该gene的可信度

sscore: gene的转录起始位置分值, 为以下三种之和: rscore, 该gene RBS motif分值; uscore: 围绕起始密码分值; tscore: 起始密码子类型分值

```
DALY01000013.1_1_5 # 2294 # 2950 # -1 #
ID=1_5;partial=00;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp
ATGAGGGCCGATCTTCACGTTCAATTCAAGTTACTCCAATGACGGAGTCTCAACGCCTCAGCAGA
TCGTCG
ACAGAGCGATAGAGGTAGGTTTGGGATGTGTTGCGATCACGGATCACAACAGTTTCAAGGCTTA
TTATGA
CGTGAAGGACAACGGAAAGATAATCATCATACCGGGCGAGGAGGTCTCCTCGAAAGAGGGACA
TATCCTT
GCCTACGGCATCAATAAGGAGATCCCGCGCGGAATGAGTATCCAGGATACCATTGATGCGATCC
ACGAGG
CCGGCGGAGTGGCCTTTGCCGCCCATCCGTACAGGTGGTGGTCCGGTCTCGGGGAGAAAAAT
ACCTTACA
GTATGATTTTGACGGTACCGAAGCAAGGAATGCAAGGTCCGTACCGCGTGCCAACAGAAGGTC
CGAGGCA
TTGGCAAAGAAGATCGGAAAACCGATATCTGCCGGCAGCGATGCCCACTCGCCTCCGAGGATC
GGTTCCG
GTTCCGGTAGACCTTCCCGACGGTCTTACCACCTGGCAGGAGGTATTAGATCACATCATGAATCAT
GATGT
CAAGGTAGACAGTACCAGCCGCGGAAGGACAGCATCGCTGAGATACGGTATCAAATCCATCGGT
CAATGG ATGTTCCGCGGTTTCAGGAAGATGTAA
```

## augustus输出格式简述

- gff3格式[<http://gmod.org/wiki/GFF3>]

```
##gff-version 3
ctg123 . exon 1300 1500 . + . ID=exon00001
ctg123 . exon 1050 1500 . + . ID=exon00002
ctg123 . exon 3000 3902 . + . ID=exon00003
ctg123 . exon 5000 5500 . + . ID=exon00004
ctg123 . exon 7000 9000 . + . ID=exon00005
```

以上各列分别为：1，seqid；2，feature source；3，feature type；4/5，feature start/end；6，feature score，使用E表示序列相似度，P值表示ab initio基因预测值，.表示没有值；7，feature strand；8，feature phase，0表示下一个密码子从该区域第一个碱基开始，1表示从第二个碱基开始，2表示从第三个碱基开始，.表示没有；9，feature attributes，格式tag=value，各attributes以分号隔开

## Usage

1. Predict genes ab initio

仅输入target genome，不使用evidence(hints)，例如果蝇chr2R, 7000,001-7500,000

--species=SPECIES，指定查询序列物种，详细支持物种见：augustus --species=help

--strand=both/forward/backward, 指定预测链, 默认both

--genemodel=partial/intronless/complete/ateastone/exactlyone, 预测基因模式, partial: 允许预测处于序列两端的不完整基因(默认), **intronless**, 仅预测单个外显子基因, 例如在原核和一些真核生物中; complete, 仅预测完整的基因; atleastone: 预测至少一个完整的基因; exactlyone, 预测精确的一个完整基因

--singlestrand=true, 每条链独立预测基因, 允许一条链相反方向出现重叠基因

--hitsfile=histsfilename, 当使用该选项时, 预测将考虑hints(extrinsic information)

--predictionStart=A, --predictionEnd=B, 定义预测序列范围

--gff3=on/off, 输出gff3格式

```
augustus --gff3=on --species=fly --predictionStart=7000001 --predictionEnd=7500000 chr2R.fa > augustus.abinitio.gff3
```

## 2. Make a Custom Gene Prediction Track on the UCSC Genome Browser

检查预测文件和UCSC注释, 首先构建基因预测结构的信息, 然后输入预测文件信息

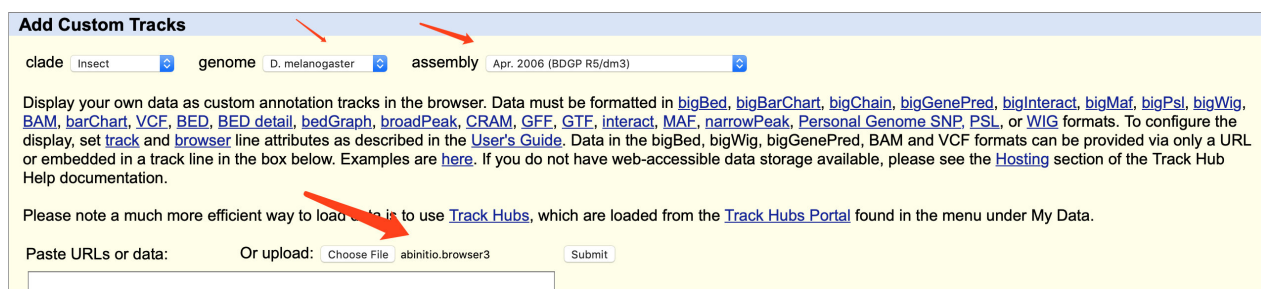
```
echo -e "browser position chr2R:7000000-7050000\n\
```

```
browser hide multiz15way bdnnpChipper\n\
```

```
track name=abinitio description=\"Augustus ab initio predictions\" db=dm3visibility=3" > abinitio.browser3
```

```
grep "AUGUSTUS\tCDS" augustus.abinitio.gff3 >> abinitio.browser3
```

UCSC custom track link[[http://genome.ucsc.edu/cgi-bin/hgCustom?hgid=729755039\\_0V2JNzjASTx6ApZ9jgUoHBUDclRq&clade=&org=D.+melanogaster&db=dm3&hgct\\_do\\_add=1](http://genome.ucsc.edu/cgi-bin/hgCustom?hgid=729755039_0V2JNzjASTx6ApZ9jgUoHBUDclRq&clade=&org=D.+melanogaster&db=dm3&hgct_do_add=1)]



**Add Custom Tracks**

clade: Insect genome: D. melanogaster assembly: Apr. 2006 (BDGP R5/dm3)

Display your own data as custom annotation tracks in the browser. Data must be formatted in [bigBed](#), [bigBarChart](#), [bigChain](#), [bigGenePred](#), [bigInteract](#), [bigMaf](#), [bigPsl](#), [bigWig](#), [BAM](#), [barChart](#), [VCF](#), [BED](#), [BED detail](#), [bedGraph](#), [broadPeak](#), [CRAM](#), [GFF](#), [GTF](#), [interact](#), [MAF](#), [narrowPeak](#), [Personal Genome SNP](#), [PSL](#), or [WIG](#) formats. To configure the display, set [track](#) and [browser](#) line attributes as described in the [User's Guide](#). Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are [here](#). If you do not have web-accessible data storage available, please see the [Hosting](#) section of the Track Hub Help documentation.

Please note a much more efficient way to load data is to use [Track Hubs](#), which are loaded from the [Track Hubs Portal](#) found in the menu under My Data.

Paste URLs or data: Or upload:  abinitio.browser3

## Introduction

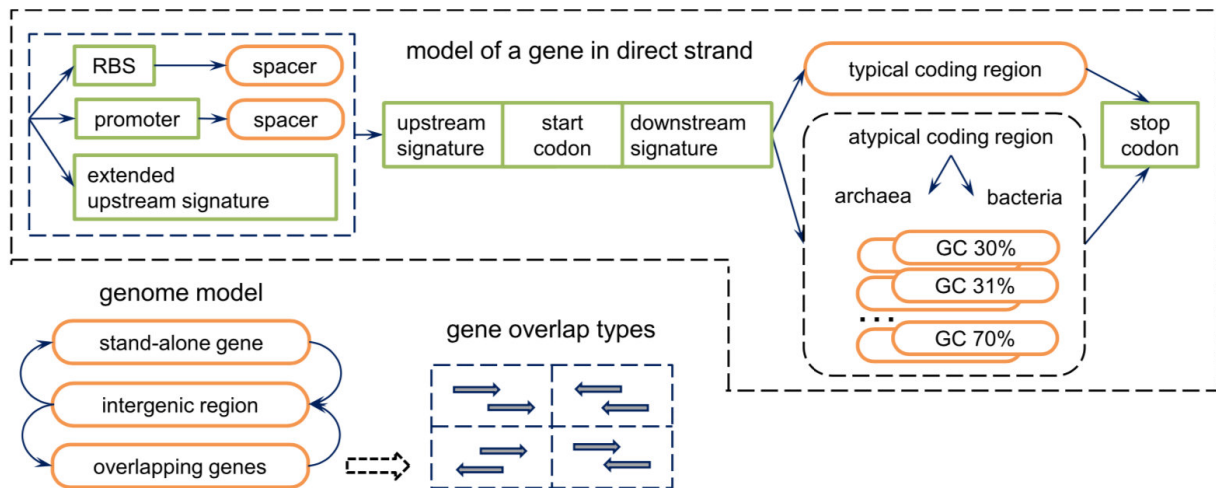
由于原核基因组RNA-seq数据几乎不能给基因预测提供额外的证据, 因此原核基因组的基因从头预测被认为是更准确的基因预测过程。

GeneMarkS-2, 根据局部基因组GC含量(GC含量变化大部分由于水平基因传递导致)选择最匹配参数, 同时采用新的motif搜索算法, LFinder, 用于检测基因起点上游调节区域的保守motif, 来预测基因起始位置。

当前流行的基因预测软件, GeneMarkS, Glimmer和Prodigal。

**Shine-Dalgarno (SD) signal has long been viewed as the dominant translation initiation signal in prokaryotes. leaderless genes, which lack 5'-untranslated regions (5'-UTR) on their mRNAs, have been shown abundant in archaea.**

原核物种之间显著的差异在于围绕基因起点序列的不同组成，这些差异反应了不同的转录和翻译机制。绝大多数原核物种不采用leaderless 转录，因此每个基因前都存在核糖体结合位点(RBS)。其他物种转录采用的为first-genes-in-operons为leaderless，这些基因的转录没有RBS位点的。基因组中，观察到的所有类型的RBS位点都处于操纵子内部基因的前面。



**Figure 1.** Principal state diagram of the generalized hidden Markov model (GHMM) of prokaryotic genomic sequence. States shown in the *top* panel were used to model a gene in the direct strand. Genes in the reverse strand were modeled by the identical set of states (with directions of transition reversed). The states modeling genes in direct and reverse strands were connected through the intergenic region state as well as the states of genes overlapping in opposite strands.

## Usage

`gms2.pl --seq SEQ --genome-type TYPE`

`--seq`, 包含fasta格式序列的输入文件

`--genome-type`, archaea, bacteria, auto

`--gcode`, 遗传密码子，默认为11，可选11或4

`--output`, 输出文件，默认为gms2.lst

`--ext`, 外部gff格式文件信息

`--fnn`, 输出预测基因的核酸序列

`--faa`, 输出预测基因的蛋白序列

`--gid`, 改变gene ID格式

```
gms2.pl --seq 38588-scaffolds.fasta --genome-type bacteria --output
38588_gms2.lst --fnn 38588_gms2.fnn --faa 38588_gms2.faa
```

