

利用 Python 进行生信分析 | 基础 + 实战

——Python for Bioinformatics



赵华男

2022 一起加油

项目三 基因注释文件处理

基因注释与 GTF 和 GFF 文件的介绍

基因注释与 GTF 和 GFF 文件的介绍

● GFF 和 GTF 是两种最常用的数据库注释格式:

- GFF 全称为 general feature format, 主要用来注释基因组
- GTF 全称为 gene transfer format, 主要用来注释基因

● 注释文件的用途:

- 在生物信息学分析中, 我们不但需要参考基因组信息 (FASTA) 和二代测序数据 (FASTQ) 来进行测序数据比对回贴
- 还需要与之对应的注释信息 (GFF, GTF) 来进行下游分析, 比如常见的 RNA-seq (transcript) 和 ChIP (gene)

● GTF 是在 GFF 的基础上发展而来:

- 本质上都是 TSV 文件 (以 TAB 制表符分隔的文本文件)
- 都是 9 列文件, 内容也比较接近
- GFF 能够包含的信息更多更全, 可以包含染色体, 基因, 转录本的信息
- 而 GTF 主要用来描述基因和转录本的信息

● 相互转化:

- 如使用 Cufflinks 软件的的 gffread 命令
- 我们自己写一个!

● 查看官方描述 (下面的链接)(GFF2 已经弃用, 特指 GFF3)

<http://www.gmod.org/wiki/GFF3>

<https://mblab.wustl.edu/GTF22.html>

Demo: GTF

File: hg38_refseq_from_ucsc.rm_XM.XR.fix_name.gtf									
1	chr1	hg38_ncbiRefSeq	exon	67092176	67093604	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
2	chr1	hg38_ncbiRefSeq	exon	67096252	67096321	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
3	chr1	hg38_ncbiRefSeq	exon	67111577	67111644	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
4	chr1	hg38_ncbiRefSeq	exon	67111577	67111644	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
5	chr1	hg38_ncbiRefSeq	exon	67113614	67113756	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
6	chr1	hg38_ncbiRefSeq	exon	67115352	67115464	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
7	chr1	hg38_ncbiRefSeq	exon	67125752	67125909	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
8	chr1	hg38_ncbiRefSeq	exon	67127166	67127257	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
9	chr1	hg38_ncbiRefSeq	exon	67131142	67131227	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
10	chr1	hg38_ncbiRefSeq	exon	67134930	67134971	0.000000	-	.	gene_id "Clorf141"; transcript_id "NR_075077.1";
11	chr1	hg38_ncbiRefSeq	stop_codon	67093580	67093582	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
12	chr1	hg38_ncbiRefSeq	CDS	67093583	67093604	0.000000	-	1	gene_id "Clorf141"; transcript_id "NM_001276352.1";
13	chr1	hg38_ncbiRefSeq	exon	67092176	67093604	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
14	chr1	hg38_ncbiRefSeq	CDS	67096252	67096321	0.000000	-	2	gene_id "Clorf141"; transcript_id "NM_001276352.1";
15	chr1	hg38_ncbiRefSeq	exon	67096252	67096321	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
16	chr1	hg38_ncbiRefSeq	CDS	67103238	67103382	0.000000	-	0	gene_id "Clorf141"; transcript_id "NM_001276352.1";
17	chr1	hg38_ncbiRefSeq	exon	67103238	67103382	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
18	chr1	hg38_ncbiRefSeq	CDS	67111577	67111644	0.000000	-	2	gene_id "Clorf141"; transcript_id "NM_001276352.1";
19	chr1	hg38_ncbiRefSeq	exon	67111577	67111644	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
20	chr1	hg38_ncbiRefSeq	CDS	67115352	67115464	0.000000	-	1	gene_id "Clorf141"; transcript_id "NM_001276352.1";
21	chr1	hg38_ncbiRefSeq	exon	67115352	67115464	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
22	chr1	hg38_ncbiRefSeq	CDS	67125752	67125909	0.000000	-	0	gene_id "Clorf141"; transcript_id "NM_001276352.1";
23	chr1	hg38_ncbiRefSeq	exon	67125752	67125909	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
24	chr1	hg38_ncbiRefSeq	CDS	67127166	67127240	0.000000	-	0	gene_id "Clorf141"; transcript_id "NM_001276352.1";
25	chr1	hg38_ncbiRefSeq	start_codon	67127238	67127240	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
26	chr1	hg38_ncbiRefSeq	exon	67127166	67127257	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
27	chr1	hg38_ncbiRefSeq	exon	67131142	67131227	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
28	chr1	hg38_ncbiRefSeq	exon	67134930	67134971	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276352.1";
29	chr1	hg38_ncbiRefSeq	stop_codon	67093005	67093007	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276351.1";
30	chr1	hg38_ncbiRefSeq	CDS	67093008	67093604	0.000000	-	0	gene_id "Clorf141"; transcript_id "NM_001276351.1";
31	chr1	hg38_ncbiRefSeq	exon	67092176	67093604	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276351.1";
32	chr1	hg38_ncbiRefSeq	CDS	67095235	67095421	0.000000	-	1	gene_id "Clorf141"; transcript_id "NM_001276351.1";
33	chr1	hg38_ncbiRefSeq	exon	67095235	67095421	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276351.1";
34	chr1	hg38_ncbiRefSeq	CDS	67096252	67096321	0.000000	-	2	gene_id "Clorf141"; transcript_id "NM_001276351.1";
35	chr1	hg38_ncbiRefSeq	exon	67096252	67096321	0.000000	-	.	gene_id "Clorf141"; transcript_id "NM_001276351.1";

Demo: another GTF

```
##description: evidence-based annotation of the human genome (GRCh38), version 29 (Ensembl 94)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2018-08-30
chr1 HAVANA gene 11869 14409 . + . gene_id "ENSG00000223972.5"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana gene
chr1 HAVANA transcript 11869 14409 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana transcript
chr1 HAVANA exon 11869 12227 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 12613 12721 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 13221 14409 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA transcript 12010 13670 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana transcript
chr1 HAVANA exon 12010 12057 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 12179 12227 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 12613 12697 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 12975 13052 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 13221 13374 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA exon 13453 13670 . + . gene_id "ENSG00000223972.5"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana exon
chr1 HAVANA gene 14404 29570 . - . gene_id "ENSG00000227232.5"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana gene
chr1 HAVANA transcript 14404 29570 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana transcript
chr1 HAVANA exon 29534 29570 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 24738 24891 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 18268 18366 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 17915 18061 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 17606 17742 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 17233 17368 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 16858 17055 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 16607 16765 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 15796 15947 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 15005 15038 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 HAVANA exon 14404 14501 . - . gene_id "ENSG00000227232.5"; transcript_id "ENST00000488147.1"; gene_type "unprocessed_pseudogene"; gene_name "WASH7P"; level 2; havana exon
chr1 ENSEMBL gene 17369 17436 . - . gene_id "ENSG00000278267.1"; gene_type "miRNA"; gene_name "MIR6859-1"; level 3;
chr1 ENSEMBL transcript 17369 17436 . - . gene_id "ENSG00000278267.1"; transcript_id "ENST00000619216.1"; gene_type "miRNA"; gene_name "MIR6859-1"; level 3;
chr1 ENSEMBL exon 17369 17436 . - . gene_id "ENSG00000278267.1"; transcript_id "ENST00000619216.1"; gene_type "miRNA"; gene_name "MIR6859-1"; level 3;
chr1 HAVANA gene 29554 31109 . + . gene_id "ENSG00000243485.5"; gene_type "lincRNA"; gene_name "MIR1302-2HG"; level 2; tag "ncRNA_host"; havana gene
chr1 HAVANA transcript 29554 31097 . + . gene_id "ENSG00000243485.5"; transcript_id "ENST00000473358.1"; gene_type "lincRNA"; gene_name "MIR1302-2HG"; level 2; havana transcript
chr1 HAVANA exon 29554 30039 . + . gene_id "ENSG00000243485.5"; transcript_id "ENST00000473358.1"; gene_type "lincRNA"; gene_name "MIR1302-2HG"; level 2; havana exon
chr1 HAVANA exon 30564 30667 . + . gene_id "ENSG00000243485.5"; transcript_id "ENST00000473358.1"; gene_type "lincRNA"; gene_name "MIR1302-2HG"; level 2; havana exon
```

Demo: GFF3

```
##gff-version 3.1.26
##sequence-region ctg123 1 1497228
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA 1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA 1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon 1300 1500 . + . ID=exon00001;Parent=mRNA00003
ctg123 . exon 1050 1500 . + . ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon 3000 3902 . + . ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon 5000 5500 . + . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon 7000 9000 . + . ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS 1201 1500 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS 3000 3902 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS 5000 5500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS 7000 7600 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS 1201 1500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS 5000 5500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS 7000 7600 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS 3301 3902 . + 0 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
```

Info: 以 GTF 为例

● 1. seq_id

- 序列的编号, 一般为 chr 或者 scaffold 编号, 每条染色体拥有一个唯一的 ID。

● 2. source

- 注释的来源, 代表基因结构的来源, 可以是数据库的名称, 比如来自 RefSeq 数据库, 也可以是软件的名称, 比如用 GeneScan 软件预测得到, 当然, 也可以为空, 用 “.” 点号填充。

● 3. type

- 代表区间对应的特征类型, 在 GTF 中, 常见的类型如下:
 - Gene, cDNA, mRNA, 5UTR, 3UTR, exon, CDS, start_codon, stop_codon

● 4. start

- 该基因或转录本在参考序列上的起始位置。

● 5. end

- 该基因或转录本在参考序列上的终止位置。

● 6. score

- 得分, 软件提供了统计值, 是注释信息可能性的说明, 可以是序列相似性比对时的 E-values 值或者基因预测是的 P-values 值
- “.” 表示为空。

● 7. strand

- 代表正负链的信息, + 表示正链, - 表示负链, ? 表示不清楚正负链的信息
- 当正负链信息没有意义时, 可以用 “.” 填充。

Info: 以 GTF 为例

● 8. phase

- 仅对注释类型为“CDS”有效, 表示起始编码的位置, 有效值为 0、1、2
 - 对于编码蛋白质的 CDS 来说, 本列指定下一个密码子开始的位置。每 3 个核苷酸翻译一个氨基酸, 从 0 开始, CDS 的起始位置, 除以 3, 余数就是这个值, 表示到达下一个密码子需要跳过的碱基数。
 - 0 表示该编码框的第一个密码子第一个碱基位于其 5' 末端
 - 1 表示该编码框的第一个密码子的第一个碱基位于该编码区外
 - 2 表示该编码框的第一个密码子的第一、二个碱基位于该编码区外
- 如果 **Feature** 为 CDS 时, 必须指明具体值!

● 9. attributes (更正!)

- 一个包含众多属性的列表
 - GFF3: 格式为“标签 = 值” | “(tag=value)”, 如果一个序列元件没有 Parent 属性, 说明他的父元件就是 scaffold 或者 chromosome
 - GTF: 格式为“标签值;” | “tag value;”, 每个特征之后都要有分号 (包括最后一个特征)
 - GTF: 其内容必须包括 **gene_id** 和 **transcript_id**, 其 value 可以为空 “”

GTF 和 GFF 文件操作

● ensembl 示例文件下载

- <https://www.ensembl.org/>
- All genomes (Search 下方)
- 选择 Human
- Gene annotation
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

Index of /pub/release-107/gtf/homo_sapiens

Name	Last modified	Size	Description
 Parent Directory			-
 CHECKSUMS	2022-06-05 20:13	225	
 Homo_sapiens.GRCh38.107.abinitio.gtf.gz	2022-06-04 11:52	3.3M	
 Homo_sapiens.GRCh38.107.chr.gtf.gz	2022-06-04 11:45	50M	
 Homo_sapiens.GRCh38.107.chr_patch_hapl_scaff.gtf.gz	2022-06-04 11:49	54M	
 Homo_sapiens.GRCh38.107.gtf.gz	2022-06-04 11:45	50M	
 README	2022-06-04 11:50	9.9K	

GTF 和 GFF 文件操作

Demo: GFF3 feature

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv(
5     'data/Homo_sapiens.GRCh38.107.gff3.gz',
6     sep="\t",
7     comment="#",
8     names=Annotation.REQUIRED_COLUMNS,
9     skipinitialspace=True,
10    skip_blank_lines=True,
11    on_bad_lines='skip',
12    # chunksize=chunksize,
13    engine="c",
14    dtype={
15        "start": np.int64,
16        "end": np.int64,
17        "score": np.float32,
18        "seqname": str,
19    },
20    na_values=".",
21 )
```

GTF 和 GFF 文件操作

Demo: GFF3 feature

```
1 df.feature.unique()
2 array(['chromosome', 'biological_region', 'pseudogene', 'lnc_RNA', 'exon',
3        'pseudogenic_transcript', 'ncRNA_gene', 'miRNA', 'gene', 'mRNA',
4        'five_prime_UTR', 'CDS', 'three_prime_UTR', 'snRNA', 'ncRNA',
5        'unconfirmed_transcript', 'snoRNA', 'scRNA', 'rRNA',
6        'V_gene_segment', 'D_gene_segment', 'J_gene_segment',
7        'C_gene_segment', 'transcript', 'scaffold', 'tRNA'], dtype=object)
8
9 # feature required:
10 # - "CDS",
11 # - "start_codon" # 简化, 先不考虑, 其实是CDS前三个碱基
12 # - "stop_codon" # 简化, 先不考虑
13
14 # optional:
15 # - "5UTR",
16 # - "3UTR",
17 # - "*RNA"
```

练习

```
1 - 仅在必要时才对 self 或 cls 注释(实战课三当中进行讲解)
2 - cls + @classmethod
3 - GFF与GTF文件进行转换
4
5 - 统计各条染色体的基因密度
6 - 获得基因的的TSS, TES及启动子坐标 (GFF3)
7 - 计算全基因组可转录区域长度及所占基因组比例 (GFF3, GTF)
8 - 计算基因的转录长度、外显子数目及翻译区长度 (GFF3)
9     - 计算基因的转录长度
10     - 统计平均长度, 中位数长度
11     - 计算基因的外显子(exon)个数
12     - 统计平均个数, 中位数个数
13     - 统计最多exon的基因, 最少exon的gene
14 - 程序 CLI 化
15     - sys.argv
16     - argparse
```

程序 CLI 化

程序 CLI 化

Q & A!

Thank you!