
1 FASTA 格式

FASTA 格式（又称为 Pearson 格式），是一种基于文本用于表示核苷酸序列或氨基酸序列的格式。在这种格式中碱基对或氨基酸用单个字母来编码，且允许在序列前添加序列名及注释。序列文件的第一行是由大于号">"或分号";"打头的任意文字说明（习惯常用">"作为起始），用于序列标记。从第二行开始为序列本身，只允许使用既定的核苷酸或氨基酸编码符号。通常核苷酸符号大小写均可，而氨基酸常用大写字母。如：

```
>scaffold1 35.9
```

```
AACTCCAAATGTTTTACATCCTTTTTTTATCCATAATATATAATCAACTGATATACAAAATGAAAAAATACTACCTACATTTTTATTAGGC  
TTATTTTATTAAAATAAGGTTGGTGTGTGGAAATAGCCATTCT.....
```

2 FASTQ 格式

FASTQ 是一种存储了生物序列（通常是核酸序列）以及相应的质量评价的文本格式。

FASTQ 格式下，每个序列共有 4 行，第 1 行是 @ 序列 ID，包括 index 序列及 read1 或 read2 标志，由测序仪产生；第 2 行是碱基序列，大写“ACTGN”；第三行是“+”，省略了序列 ID；第 4 行是序列对应的测序质量值序列，每个字母对应第 2 行每个碱基，第四行每个字母对应的 ASCII 值减去 33，即为该碱基的测序质量值，比如 I 对应的 ASCII 十进制值为 73，那么其对应的碱基质量值是 40。

[illegible]

3 GFF 格式

GFF 格式是 Sanger 研究所定义，是一种简单的、方便的对于 DNA、RNA 以及蛋白质序列的特征进行描述的一种数据格式，比如序列的哪里到哪里是基因，已经成为序列注释的通用格式，比如基因组的基因预测，许多软件都支持输入或者输出 GFF 格式。目前格式定义的最新版本是版本 3。GFF 格式举例如下：

```
Scaffold1      GeneMark      gene      1007      1627      .      +      .      ID=artGM000001;Name=artGM000001;
Scaffold1      GeneMark      mRNA      1007      1627      .      +      .      ID=artGM000001;Name=artGM000001;Parent=artGM000001;
Scaffold1      GeneMark      CDS       1007      1627      .      +      0      Parent=artGM000001;
Scaffold1      GeneMark      gene      1590      2651      .      -      .      ID=artGM000002;Name=artGM000002;
Scaffold1      GeneMark      mRNA      1590      2651      .      -      .      ID=artGM000002;Name=artGM000002;Parent=artGM000002;
Scaffold1      GeneMark      CDS       1590      2651      .      -      0      Parent=artGM000002;
```

文件格式说明见下表：

列数	说明
1	“seqid”序列的编号，编号的有效字符有[a-zA-Z0-9.:^x!+_-]
2	“source”注释信息的来源，比如”Genescan”、”Genbank” 等，可以为空，为空用”.”点号代替
3	“type”注释信息的类型，比如 Gene、cDNA、mRNA 等，或者是 SO 对应的编号
4	“start”起始位置
5	“end”终止位置
6	“score”得分，数字，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 值或者基因预测时的 P-values 值。”.”表示为空
7	“strand”序列的方向， +表示正义链, -反义链 ,?表示未知
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。 “attributes”以多个键值对组成的注释信息描述，键与值之间用“=”，不同的键值用“;”隔开，一个键可以有多个值，不同值用“,”分割。注意如果描述中包括 tab 键以及“=;”，要用 URL 转义规则进行转义，如 tab 键用 %09 代替。键是区分大小写的，以大写字母开头的键是预先定义好的，在后面可能被其他注释信息所调用。

4 m8 格式

m8 格式为列表格式的 BLAST 比对结果。m8 格式举例如下：

artGM000002	YP_925440.1	61.54	338	126	2	3	340	11	344	1e-127	377
artGM000003	YP_925441.1	81.40	688	127	1	30	716	7	694	0.0	1122
artGM000004	YP_925442.1	81.46	205	38	0	1	205	1	205	5e-123	353
artGM000007	YP_925444.1	85.33	259	38	0	1	259	1	259	1e-170	478
artGM000011	YP_925448.1	63.07	287	98	3	3	288	5	284	6e-106	317
artGM000012	YP_925449.1	50.31	322	156	2	27	344	36	357	3e-93	290

文件内容说明如下：

列数	说明
1	目标核酸或氨基酸序列的 ID，编号的有效字符有[a-zA-Z0-9.:^x!+_?~]。
2	数据库序列的 ID。
3	目标核酸或氨基酸序列与数据库序列比对的 Identity 值。
4	目标核酸或氨基酸序列与数据库序列比对的长度。
5	目标核酸或氨基酸序列与数据库序列比对区域的比对错配数。
6	目标核酸或氨基酸序列与数据库序列比对区域的比对空位数。
7	目标核酸或氨基酸序列的比对起始坐标。
8	目标核酸或氨基酸序列的比对终止坐标。
9	数据库序列的比对起始坐标。
10	数据库序列的比对终止坐标。
11	目标核酸或氨基酸序列与数据库序列比对的期望值。
12	目标核酸或氨基酸序列与数据库序列比对的比对得分。