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:

```
>gene_2
ATGCTGGAGAATCAGGGATTGAAGAAAAGGCAGCTCTTTTCCGCGCAGGTGGATTTGAGT
AAATTTAATGATAAAGATTTTGACCAGGCGACAGAGGATGAGAAGAAGCAGCAGGATGTT
ATGGGGGAGTCAACGACTTTCTTCAAAGACGGCATGCGGCGTTTGCGCAAGAACCCTCTG
GCGATGGGGAGTATTGTGGTGCTGGTGTTG
```

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。

```
@A00821:894:HL2JKDSX2:1:1101:4227:1016 1:N:0:ACGAATTGCC+CGTTATTCGG
GNTCAGTTTGACTATATCATTTTGGACTGCCCGGCAGGGATTGAACAGGGCTTTCAGAATGCCATCGCCGGCG
CGGACAGGGCGCTGGTGGTAACAACACCGGAAGTCTCGGCTATCCGGGATGCAGACAGGATCATCGGTCTGTTAG
AA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FF
```

m8 格式

m8 格式为列表格式的 BLAST 比对结果。m8 格式举例如下：

HC1_gene_188	ccel:CCDG5_1826	78.7	178	38	0	1	178	301	478	1.92e-90	293
HC1_gene_191	bdo:EL88_22335	56.8	491	204	5	1	489	1	485	5.71e-184	530
HC1_gene_192	bdo:EL88_00840	73.5	381	101	0	1	381	1	381	8.07e-199	559
HC1_gene_209	dun:FDZ78_14020	37.5	283	162	6	30	305	13	287	1.13e-51	178
HC1_gene_212	rix:RO1_09980	92.1	89	7	0	1	89	77	165	3.59e-48	156
HC1_gene_213	rix:RO1_09990	89.2	65	7	0	1	65	1	65	5.28e-31	108
HC1_gene_214	rix:RO1_10000	93.2	118	8	0	1	118	1	118	6.87e-67	203
HC1_gene_223	csr:Cspa_c41080	45.2	270	139	1	7	276	219	479	6.45e-77	248
HC1_gene_224	rus:RBI_I02016	77.0	517	112	2	1	516	1	511	1.19e-298	822
HC1_gene_226	gca:Galf_0970	37.6	258	145	6	4	260	2	244	1.92e-45	160
HC1_gene_233	vgu:HYG85_07040	72.2	162	45	0	1	162	40	201	4.80e-76	235
HC1_gene_241	csr:Cspa_c29540	67.5	166	54	0	13	178	1	166	5.72e-80	245

文件内容说明如下：

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。