**DNA各种序列格式介绍**

2011年12月18日 ⁄ [Genomics](http://www.plob.org/category/genomics) ⁄ 字号 [小](javascript:doZoom(12)) [中](javascript:doZoom(13)) [大](javascript:doZoom(15)) ⁄ [暂无评论](http://www.plob.org/2011/12/18/1193.html#respond) ⁄ 阅读 510 次 [[点击加入在线收藏夹]](javascript:void(0);)

**1.**[**Plain**](http://www.plob.org/tag/plain)**格式**

A sequence in plain format may contain only [IUPAC](http://www.plob.org/tag/iupac) characters and spaces (no numbers!).  
Note: A file in plain sequence format may only contain one sequence, while most other formats accept several sequences in one file.  
An example sequence in plain format is:  
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGC  
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGA  
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGC  
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCG  
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCAT  
TTTAATTACAGACCTGAA

[Plain](http://www.plob.org/tag/plain) sequence序列格式，只含有[IUPAC](http://www.plob.org/tag/iupac" \o "View all posts in IUPAC" \t "_blank)字符和空格，不含有数字，并且一个[Plain](http://www.plob.org/tag/plain" \o "View all posts in Plain" \t "_blank)格式的文件只能含有一条序列。

**2.**[**EMBL**](http://www.plob.org/tag/embl)**格式**

A sequence file in [EMBL](http://www.plob.org/tag/embl) format can contain several sequences.  
One sequence entry starts with an identifier line (“ID”), followed by further annotation lines. The start of the sequence is marked by a line starting with “SQ” and the end of the sequence is marked by two slashes (“//”).  
An example sequence in [EMBL](http://www.plob.org/tag/embl) format is:  
ID   AB000263 standard; RNA; PRI; 368 BP.  
XX  
AC   AB000263;  
XX  
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.  
XX  
SQ   Sequence 368 BP;  
acaagat[gcc](http://www.plob.org/tag/gcc) attgtccccc g[gcc](http://www.plob.org/tag/gcc)tcctgc tgctgctgct ctccggg[gcc](http://www.plob.org/tag/gcc) acg[gcc](http://www.plob.org/tag/gcc)accg        60  
ct[gcc](http://www.plob.org/tag/gcc)ct[gcc](http://www.plob.org/tag/gcc) cctggagggt g[gcc](http://www.plob.org/tag/gcc)ccaccg [gcc](http://www.plob.org/tag/gcc)gagacag cgagcatatg caggaagcgg       120  
caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180  
ag[gcc](http://www.plob.org/tag/gcc)agtgc cgg[gcc](http://www.plob.org/tag/gcc)cctc ataggagagg aagctcggga ggtg[gcc](http://www.plob.org/tag/gcc)agg cggcaggaag       240  
gcgcaccccc ccagcaatcc gcgc[gcc](http://www.plob.org/tag/gcc)ggg acagaat[gcc](http://www.plob.org/tag/gcc) ctgcaggaac ttcttctgga       300  
agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360  
gacctgaa                                         368  
//

[EMBL](http://www.plob.org/tag/embl)格式文件可以包含多条序列，每个序列条目都以”ID”开始，紧跟一些注释信息，序列的开始标记为”SQ”，结束标记为”//”。

**3.FASTA格式**

A sequence file in FASTA format can contain several sequences.  
Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (“>”) symbol in the first column.  
An example sequence in FASTA format is:  
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA  
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCC  
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAA  
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCC  
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGC  
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATG  
TTTAATTACAGACCTGAA

FASTA格式文件可以包含多条序列，每条序列之前都有以”>”开始的一行，该行包含一些序列的描述信息。

**4.GCG格式**

A sequence file in GCG format contains exactly one sequence, begins with annotation lines and the start of the sequence is marked by a line ending with two dot (“..”) characters. This line also contains the sequence identifier, the sequence length and a checksum. This format should only be used if the file was created with the GCG package.  
An example sequence in GCG format is:  
ID   AB000263 standard; RNA; PRI; 368 BP.  
XX  
AC   AB000263;  
XX  
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.  
XX  
SQ   Sequence 368 BP;  
AB000263  Length: 368  Check: 4514  ..  
1  acaagat[gcc](http://www.plob.org/tag/gcc) attgtccccc g[gcc](http://www.plob.org/tag/gcc)tcctgc tgctgctgct ctccggg[gcc](http://www.plob.org/tag/gcc) acg[gcc](http://www.plob.org/tag/gcc)accg  
61  ct[gcc](http://www.plob.org/tag/gcc)ct[gcc](http://www.plob.org/tag/gcc) cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg  
121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc  
181  aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag  
241  gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga  
301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca  
361  gacctgaa

GCG格式文件只含有一条序列，以一些注释信息行开始，序列以”..”行开始，该行还包含序列的标识，以及长度等。

**5.**[**GenBank**](http://www.plob.org/tag/genbank)**格式**

A sequence file in [GenBank](http://www.plob.org/tag/genbank) format can contain several sequences.  
One sequence in [GenBank](http://www.plob.org/tag/genbank) format starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing “OR[IG](http://www.plob.org/tag/ig)IN” and the end of the sequence is marked by two slashes (“//”).  
An example sequence in [GenBank](http://www.plob.org/tag/genbank) format is:  
LOCUS       AB000263                 368 bp    mRNA    linear   PRI 05-FEB-1999  
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.  
ACCESSION   AB000263  
OR[IG](http://www.plob.org/tag/ig)IN  
1     acaagatgcc  attgtccccc   ggcctcctgc tgctgctgct ctccggggcc acggccaccg  
61   ctgccctgcc   cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg  
121 caggaataag  gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc  
181 aggccagtgc  cgggcccctc  ataggagagg aagctcggga ggtggccagg cggcaggaag  
241 gcgcaccccc  ccagcaatcc  gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga  
301 agaccttctcc  ctcctgcaaa  taaaacctca  cccatgaatg ctcacgcaag tttaattaca  
361 gacctgaa  
//

[GenBank](http://www.plob.org/tag/genbank)格式文件可以包含多个序列，每个序列条目都以”LOCUS”开始，紧跟多行注释信息，序列开始标记为”OR[IG](http://www.plob.org/tag/ig)IN”，序列结束标记为”//”。

**6.IG格式**

A sequence file in IG format can contain several sequences, each consisting of a number of comment lines that must begin with a semicolon (“;”), a line with the sequence name (it may not contain spaces!) and the sequence itself terminated with the termination character ’1′ for linear or ’2′ for circular sequences.  
An example sequence in IG format is:  
; comment  
; comment  
AB000263  
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCG  
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAG  
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTC  
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGG  
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATG  
TTTAATTACAGACCTGAA1

IG格式序列文件可以包含多个序列，每个序列条目都以多个comment行开始，且comment行以”;”开始，comment行下面是包含序列名称的一行，序列以数字1结束，第2条序列以2结束，以此类推。

**7.**[**IUPAC**](http://www.plob.org/tag/iupac)**字符**

To represent ambiguity in [DNA](http://www.plob.org/tag/dna) sequences the following letters can be used (following the rules of the International Union of Pure and Applied Chemistry ([IUPAC](http://www.plob.org/tag/iupac))):  
A = adenine  
C = cytosine  
G = guanine  
T = thymine  
U = uracil  
R = G A (purine)  
Y = T C (pyrimidine)  
K = G T (keto)  
M = A C (amino)  
S = G C  
W = A T  
B = G T C  
D = G A T  
H = A C T  
V = G C A  
N = A G C T (any)