

# Biological Sequence Data Formats

Here we present three standard formats in which biological sequence data (DNA, RNA and protein) can be stored and presented.

**Raw Sequence:** Data without description.

**FASTA Format:** One line of description, then sequence.

**GenBank Record:** Lots of detailed description about the sequence.

## **Raw Sequence: DNA**

In the text box below, is an example of “raw” DNA sequence. Just the four different nucleotides of one particular DNA strand. No information on the gene or the organisms the sequence came from. (See “Biology in the Computer”.)

```
ATGAGTAATTCACAGCTGAGGACAAGGCTGCTATCACTAGCCTGTGGGGCAAGGTGAATGTGGAAGATG  
CTGGGGGAGAAACCCTGGGAAGGCTCCTGGTTGTACCCATGGACCCAGAGGTTCTCGACAGCTTGG  
AAGCCTGTCCTCTCCCTGCCATCATGGCAACCCCCAAAGTCAAGGCGCATGGCGTGAAGGTGCTGACT  
TCCTTGGGAGAAAGCTATAAGAACCTTGATGATCTCAAGGGCACCTTGGCCAGCTGAGTGAGCTGCAC  
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGCTCCTGGAAATGTGCTGGTACTGTTTGGCAAT  
CCTTCATGGCAAAGAATTCACCCCTGAGGTGCAGGCTCCTGGCAGAAGATGGTGGCTGGAGTGGCCAGT  
GCCTTGGCCTCCAGATACCACTGA
```

## **Raw Sequence: Protein**

Below is an example of a “raw” protein sequence. The letters indicate one of the twenty different amino acids and the order tells how they are put together.

```
MSNFTAEDKAAITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGSLSSPSAIMGNPKVKAHGVKVL  
SLGEAIKNLDDLKGTFGQLSELHCDKLHVDPENFRLLGNVLTVLAILHGKEFTPEVQASWQKMVAGVAS  
ALASRYH
```

## FASTA Format

The fasta format (originally created for a program called, you guessed it, “FASTA”) is a ubiquitous format in bioinformatics and is accepted as input to many bioinformatics analysis tools. It is almost as simple as the raw format, but has a **Title Line** that provides some information about the sequence.

FASTA formats always have a title line, and it always begins with a “>” and ends with a return character.

### FASTA Format: DNA

Below is a FASTA file for the DNA sequence that codes for the G-gamma-globin protein of a spider monkey, *Ateles geoffroyi*.

```
> Ateles geoffroyi G-gamma-globin gene, complete cds |
ATGAGTAATTCACAGCTGAGGACAAGGCTGCTATCACTAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGGGGAGAAACCTGGGAAGGCTCCTGGTTGTACCCATGGACCCAGAGGTTCTCGACAGCTTGG
AAGCCTGTCCTCTCCCTCTGCCATCATGGCAACCCCAAAGTCAAGGGCGATGGCGTGAAGGTGCTGACT
TCCTTGGAGAACGCTATAAACGACCTTGATGATCTCAAGGGCACCTTGGCCAGCTGAGTGAGCTGCAC
GTGACAAGCTGCATGTGGATCCTGAGAACCTCAGGCTCCTGGAAATGTGCTGGTGAATGTTGGCAAT
CCTCATGGCAAAGAATTCACCCCTGAGGTGCAGGCTCCTGGCAGAAGATGGTGGCTGGAGTGGCCAGT
GCCTTGGCCTCCAGATACCACTGA
```

## FASTA Format: Protein

Below is a fasta file for the Protein sequence for the G-gamma-globin protein of a spider monkey, *Ateles geoffroyi*. This is the FASTA sequence record from GenBank, a major database of biological sequence information. The codes at the beginning of the title are tracking identifiers used by GenBank to organize and find sequences in the database.



```
>gi|342383|gb|AAA36926.1| G-gamma-globin [Ateles geoffroyi]
MSNFTaedkAAItSLWGKVNVedAGGETLGRLVVYPWTQRFFDSFGSLSSPSAIMGNPKVKAHGVKVLT
SLGEAIKNLDDLKGTFGQLSELHCDKLHVDPENFRLLGNVLTVLAILHGKEFTPEVQASWQKMVAGVAS
ALASRYH
```

## FASTA Format: Make up your own titles

You do not have to have complicated titles. It is easy to make up your own titles. For example:

```
> Seq1
CCCTAAACCTAAACCTAAACCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAA
```

**WARNING:** Some programs have difficulty with titles that are too long, include spaces or non-letter or number characters. Avoid (1) Names longer than 15 character; (2) Spaces; and (3) Characters other than letters or numbers.

## FASTA Format: Multiple Entries

Sometimes you need to input many sequences at the same time to a program, such as a *multiple sequence alignment* program. This is easy in FASTA format – see below. (Note: These sequences are all the same length, but this does not have to be the case.)

```
> HumanGlobin
CCCTAAACCTAAACCTAAACCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAA
ATCTTAAATCCACCCCTAAACCTAAACCTAAACCTCTGAATCCTTAATCCCTAAAT
> MonkeyGlobin
GTATATAATGATAATTATCGTTTATGTAATTGCTTATTGTTGTGTAGATT
TTTGAGGTCAATACAAATCCTATTCCTGGTTCTTCCTTCACTTAGCTATGGA
> HorseGlobin
ATTTGTTATATTGGATAACAAGCTTGCTACGATCTACATTGGGAATGTGAGTCTT
GGGTTGGTTATCTCAAGAATCTTATTAAATTGTTGGACTGTTATGTTGGACATT
```

# GenBank Record

The GenBank format is an example of a data-rich format. It is used by The National Center for Biotechnology Information (NCBI) and each record is given a unique identification code. (Actually more than one.) The full biological sequence of the record is always at the end of the record. To the right is the GenBank record for the Spider Monkey globin gene.

Read below for more details on the types of information in a GenBank file.

**LOCUS** MNKHGBGGAG 444 bp DNA linear PRI 07-JUN-1994  
**DEFINITION** A.geoffroyi G-gamma-globin gene, complete cds.  
**ACCESSION** M36773 REGION: join(195..286,410..632,1422..1550)  
**VERSION** M36773.1 GI:342382  
**KEYWORDS** G-gamma globin; gamma-globin; hemoglobin.  
**SOURCE** Ateles geoffroyi (black-handed spider monkey)  
**ORGANISM** Ateles geoffroyi  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
Platyrrhini; Atelidae; Atelinae; Ateles.  
**REFERENCE** 1 (bases 1 to 444)  
**AUTHORS** Giebel,L.B., van Santen,V.L., Slichter,J.L. and Spritz,R.A.  
**TITLE** Nucleotide sequence, evolution, and expression of the fetal globin  
gene of the spider monkey Ateles geoffroyi  
**JOURNAL** Proc. Natl. Acad. Sci. U.S.A. 82 (20), 6985-6989 (1985)  
**PUBMED** 2413451  
**COMMENT** Original source text: Ateles geoffroyi skin DNA.  
**FEATURES**  
source 1..444  
/expression="Ateles geoffroyi"  
/mol\_type="genomic DNA"  
/db\_xref="taxon:9509"  
/cell\_type="fibroblast"  
/tissue\_type="skin"  
exon <1..92  
/note="G-gamma-globin; putative"  
/number=1  
CDS 1..444  
/note="G-gamma-globin"  
/codon\_start=1  
/protein\_id="AAA36926.1"  
/db\_xref="GI:342383"  
/translation="MSNPTAEDKAAITSLWGKVNVVEDAGGETLGRLLVVYPWTQRFDFD  
SFGSLSSPSAIMGNPKVKAHGKVVLTSLEAIFKNLDDLKGTFCQLSELHCDKLHVDPDE  
NFRLLGNVLTVLAILHGKEFTPEVQASWQKMVAGVASALASRYH"  
exon 93..315  
/note="G-gamma-globin"  
/number=2  
exon 316..>444  
/note="G-gamma-globin; putative"  
/number=3  
**ORIGIN**  
1 atgagtaatt tcacagctga ggacaaggct gctatcacta gcctgtgggg caagggtgaat  
61 gtggaaatgt ctgggggaga aaccctggga aggctctgg ttgtgtaccc atggaccgg  
121 aggttcttcg acagcttgg aagcctgtcc ttcctcttcg ccatcatggg caacccccaaa  
181 gtcaaggcgc atggcgtgaa ggtgctgact tccttggag aagctataaa gaaccttgat  
241 gatctcaagg gcacccttgg ccagctgagt gagetgcact gtgacaagct gcatgtggat  
301 cctgagaact tcaggttccctt gggaaatgtg ctggtgactg ttttggcaat ctttcatggc  
361 aaagaattca cccctggatgt gcaggcttcc tggcagaaga tggtgtggctgg agtggccagt  
421 gccttggctt ccagatcca ctga  
//

## GenBank Record: Background on the sequence

The beginning of the GenBank file contains background information such as the source of the biological molecule (what organism) and the scientists who discovered the sequence.

LOCUS	MNKHGBGGAG	On the left side of the file are keywords indicating the details present in the file.
DEFINITION	A. <i>geoffroyi</i> G-gamma-globin gene, complete cds.	The accession number and the GI (GenBank Identification) number are unique to this record.
ACCESSION	M36773	
VERSION	M36773.1	
KEYWORDS	G-gamma globin; gamma-globin; hemoglobin.	
SOURCE	Ateles <i>geoffroyi</i> (black-handed spider monkey)	The organism, or other source, of the biological sequence.
ORGANISM	Ateles <i>geoffroyi</i> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Platyrrhini; Atelidae; Atelinae; Ateles.	
REFERENCE	1 (bases 1 to 1705)	
AUTHORS	Giebel,L.B., van Santen,V.L., Slichtom,J.L. ar	
TITLE	Nucleotide sequence, evolution, and expression gene of the spider monkey <i>Ateles geoffroyi</i>	Information on the discoverers of the sequence, publications with the sequence, and more.
JOURNAL	Proc. Natl. Acad. Sci. U.S.A. 82 (20), 6985-6989 (1985)	
PUBMED	2413451	
COMMENT	Original source text: <i>Ateles geoffroyi</i> skin DNA.	
FEATURES	Location/Qualifiers	

## GenBank Record: The sequence

At the end of the file is the biological sequence. In this case it is a DNA sequence, but it may also be RNA or protein.

```
|ORIGIN
  1 ccatqqggtg gccagccttq ccttqaccaa tagctttqac aaggcaacct tqacccaatag
  61 tottagagta tcggggtqagg cccggggggcc qgtqqqtqgc tagggatqaa qaataaaaagg
121 aaggcacccctc catcagttcc acataactcgc tctgaaaacgt ctgagattat caataagctc
181 cttgtccaga cqccatc actagcctgt
241 qqqqcaaggta gaatcq gctctqgtqa
301 ccaggqacqag ggaggc tgcctctcaq
361 gatttqttqgc accttq gg ctcctqgttg
421 tgtacccatq gaccctt ccctctqcca
481 tcataqggcaa ccccaaaagtgc aaggcgcatg gcgtqaaagg gctgacttcc ttgggagaag
541 ctataaaqaaa ccttqatqat ctcaagggca cctttggcca gctqagtqag ctgcactqtg
601 acaaqctqca tqtggatcct gagaacttca gggtqagtcc aqqaqatatt qggggttqgga
661 gttaaqaaac ttcaqaggac tacctqggct gagaccccagt ggtaatqtt taaggqcctac
721 qgagtqccctc taaaaatcga gagggacaac tttqgcttcg agaaaaqagt tqtqaaacq
781 aggacaatqa cttttcttta ttagagtctq gtagaaaqaa ctttatcttt ccctcatttt
841 gattatctat ttaaaacatc tatctgaaaq caaggacaagt atggccatta aaaagatqca
901 ggcagaggca tatattqgct ccattccaagt ggagaacttt ggtggccaaa catatattgc
961 taaggctatt cctqtaattha gctggacaca tacaaaatgc tgcaaatqct tcattataaa
1021 cttacatcct ataattccaa atggggcaaa agtgtttctq ggggtgagaa agaataqaaa
1081 catttqtcct ggagtagatt ttttagtctg ttgcgagtgt gtgtatgtat gtgtgttttt
1141 ttgtgtgtq tqtqcgagca tqtgtttctt ttaaagttt cagcctacaa aatacaqggt
1201 ttgtqgttagc aagaagataq cttagattaa attatgccag tgactaatgc tqcaaggqga
1261 acagctacct ccatttaata cttatggcaaa atccaggctt tgaggaaqt taacataaggc
1321 ttgtattt atccagagg ccaaggctqga gcccctctqtt
1381 cactatc aactcaaca gctcctqgga aatgtqctqg
1441 tgactq attcaccctt tgagggtqcaq gcttcctqgc
1501 aqaagatqgt ggctqqaqtg gccagttccct tggccctccag ataccactqa aqcccctqcc
1561 catgatqcaq agctttcaaq gagttqgtttt attccgcaag caataaaaat aataaaaacta
1621 ttccqctcaa agatcacacg tgattqtcgt cagttatttt ttccctqtc ttccaaatat
1681 gcgaaaccaca aagggtttat gttga
//
```

The nucleotides (or amino acids in the case of protein) are numbered. The "c" is the first nucleotide of the file.

The end of the file is demarcated by two back slashes.

## GenBank Record: Feature section

In the middle of the file, the FEATURES section describe the various molecular features of the sequence and some of the biological activity.

FEATURES	Location/Qualifiers	
<u>source</u>	1..1705 /organism="Ateles geoffroyi" /mol_type="genomic DNA" /db_xref="taxon:9509" /cell_type="fibroblast" /tissue_type="skin"	The positions (length) of the sequence at the end of the file.
<u>exon</u>	141..286 /note="G-gamma-globin; putative" /number=1	Biological information. In this case the DNA was derived from spider monkey skin cells.
<u>CDS</u>	join(195..286, 410..632, 1422..155 /note="G-gamma-globin" /codon_start=1 /protein_id="AAA36926.1" /db_xref="GI:342383" /translation="MSNFTAAEKAITSLGKVNVEDAGGETLGRLLVVYPWTQRFFD SFGLSLSSPSAIMGPVKVAKHGKVVLTSLGEAIKNLDDLKGTFGQLSELHCDKLHVDP NFRLLLGNVLTVLAILHGKEFTPEVQASWQKMVAGVASALASRYH" 287..409 /note="G-gamma-globin" /number=1	CDS stands for CoCoding Sequence. Indicates if part or all of the sequence is translated into protein.
<u>intron</u>	410..632 /note="G-gamma-globin" /number=2	
<u>exon</u>	633..1421 /note="G-gamma-globin" /number=2	
<u>intron</u>	1422..1636 /note="G-gamma-globin, putative" /number=3	
		Accession/GI numbers of a different GenBank file with the protein sequence.
		The amino acids of the translated protein. Very convenient!