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”.)

```
ATGAGTAATTTCACAGCTGAGGACAAGGCTGCTATCACTAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGGGGAGAAACCCTGGGAAGGCTCCTGGTTGTGTACCCATGGACCCAGAGGTTCTTCGACAGCTTTGG
AAGCCTGTCCTCTCCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCGCATGGCGTGAAGGTGCTGACT
TCCTTGGGAGAAGCTATAAAGAACCTTGATGATCTCAAGGGCACCTTTGGCCAGCTGAGTGAGCTGCACT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGCTCCTGGGAAATGTGCTGGTGACTGTTTTGGCAAT
CCTTCATGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGGCTGGAGTGGCCAGT
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.

```
MSNFTAEDKAAITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGSLSSPSAIMGNPKVKAHGVKVLT
SLGEAIKNLDDLKGTFGQLSELHCDKLHVDPENFRLLGNVLVTVLAILHGKEFTPEVQASWQKMVAGVAS
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.

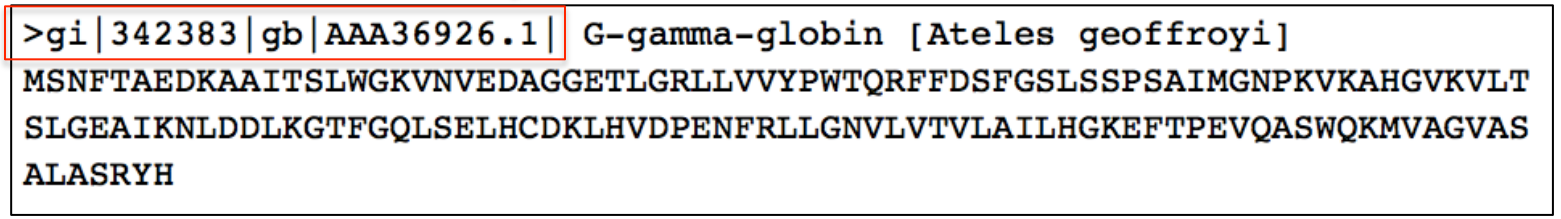
FASTA Format: DNA

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
ATGAGTAATTTCACAGCTGAGGACAAGGCTGCTATCACTAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGGGGAGAAACCCTGGGAAGGCTCCTGGTTGTGTACCCATGGACCCAGAGGTTCTTCGACAGCTTTGG
AAGCCTGTCCTCTCCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCGCATGGCGTGAAGGTGCTGACT
TCCTTGGGAGAAGCTATAAAGAACCTTGATGATCTCAAGGGCACCTTTGGCCAGCTGAGTGAGCTGCACT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAGGCTCCTGGGAAATGTGCTGGTGACTGTTTTGGCAAT
CCTTCATGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGGCTGGAGTGGCCAGT
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]
MSNFTAEDKAAITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGSLSSPSAIMGNPKVKAHGVKVLT
SLGEAIKNLDDLKGTFGQLSELHCDKLHVDPENFRLLGNVLVTVLAILHGKEFTPEVQASWQKMVAGVAS
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
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAA
```

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
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAA
ATCTTTAAATCCACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAAT
> MonkeyGlobin
GTATATAATGATAATTTTATCGTTTTTATGTAATTGCTTATTGTTGTGTGTAGATTTT
TTTGAGGTCAATACAAATCCTATTTCTTGTGGTTTTCTTTCCTTCACTTAGCTATGGA
> HorseGlobin
ATTTGTTATATTGGATACAAGCTTTGCTACGATCTACATTTGGGAATGTGAGTCTCTT
GGGTTGGTTTATCTCAAGAATCTTATTAATTGTTTGGACTGTTTATGTTTGGACATTT
```

GenBank Record

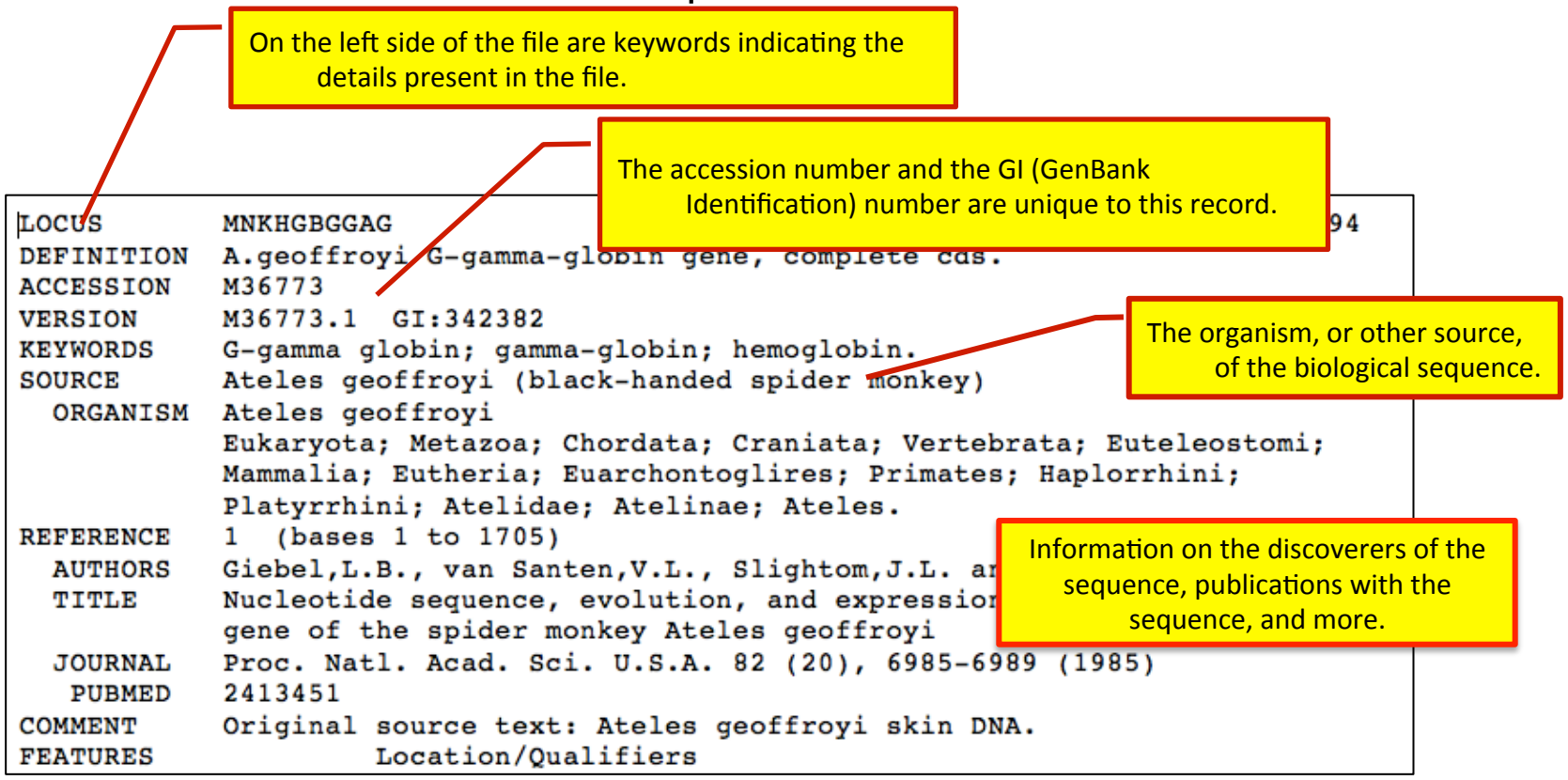
The GenBank format is an example of a data-rich format. It is used by The National Center for Biotechnology Information (NCBI) and each record is given a unique identification code. (Actually more than one.) The full biological sequence of the record is always at the end of the record. To the right is 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., Slightom,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             Location/Qualifiers
     source          1..444
                     /organism="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="MSNFTAEDKAAITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFD
                     SFGSLSSPSAIMGNPKVKAHGVKVLTSLGEAIKNLDDLKGTFGQLSELHCDKLHVDPE
                     NFRLLGNVLVTVLAILHGKEFTPEVQASWQKMVAGVASALASRYH"
     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 caaggtgaat
       61 gtggaagatg ctgggggaga aaccctggga aggctcctgg ttgtgtaccc atggacccag
      121 aggttcttcg acagctttgg aagcctgtcc tctcctctg ccatcatggg caaccccaaa
      181 gtcaaggcgc atggcgtgaa ggtgctgact tccttgggag aagctataaa gaaccttgat
      241 gatctcaagg gcacctttgg ccagctgagt gagctgcact gtgacaagct gcatgtggat
      301 cctgagaact tcaggctcct gggaaatgtg ctggtgactg ttttggcaat ccttcatggc
      361 aaagaattca cccctgaggt gcaggcttcc tggcagaaga tggtggctgg agtggccagt
      421 gccttggcct ccagatacca ctga
//
```

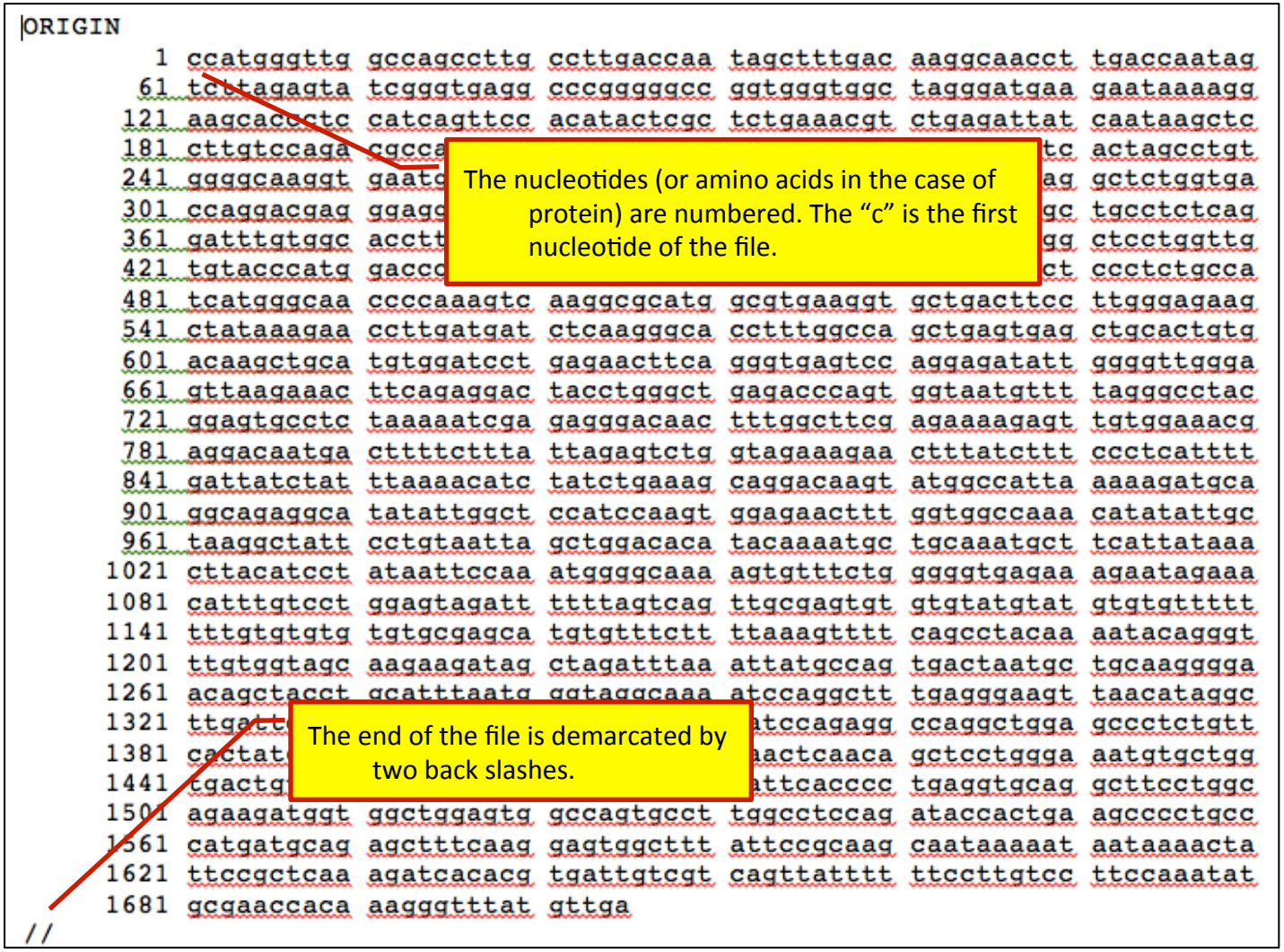
GenBank Record: Background on the sequence

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.



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.



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.