GenBank

Justin Campbell

March 7, 2000

1 What Is It?

GenBank is a free, public, online genetic sequence database. It is maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). It is available at `http://www.ncbi.nlm.nih.gov/`.

GenBank contains approximately 4,654,000,000 bases in 5,355,000 annotated sequence records as of December 1999. It is currently growing at an exponential rate, with the number of bases doubling every 14 months. A new release is made every two months. It is a part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and GenBank at NCBI. Data is exchanged between the organizations on a daily basis.

2 Searching GenBank

GenBank can be accessed through a variety of means. The different methods are accesible from `http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html`. By far, the most useful searching option is the **Entrez** system.

Entrez is a search and retrieval system that integrates information from numerous databases at NCBI, including nucleotide, protein, 3D structure, taxonomy and literature information. Entrez is available at `http://www.ncbi.nlm.nih.gov/Entrez/`. This page contains links to extensive online help. In particular, there is a tutorial for Entrez at `http://www.ncbi.nlm.nih.gov/Database/tut1.html`. From the pull-down menu, select the database you want to search, *Nucleotide* for example. Enter a few key words and then click on `Go`. The search will return a list of records which contain all of those key words. For example, a search for "University," "Washington," and "feline" returns accession number *AF192387*, the complete coding sequence for the gene FLVCR1 (submitted by researchers at the University of Washington). Clicking the link to this record will load the complete GenBank record into your browser window, in the GenBank format. The format is described in greater detail below. The tutorial, online help and FAQ all give further details on how to perform more complex searches.

3 GenBank Format

All of the GenBank data is available in the genbank directory on the NCBI ftp site at `ftp://ncbi.nlm.nih.gov/genbank/`. In particular, a thorough description of the GenBank format is available in the GenBank release notes. These are located in a file called `gbrel.txt` which resides in that directory. This includes information on downloading large collections of nucleotide sequences from their ftp site. A precise GenBank format definition is given in a postscript file in the `docs` subdirectory.

The best method for finding individual records is to use Entrez. However, complete, or nearly complete genomes are available for downloading from the `genomes/` subdirectory of the `genbank/` directory. These files are useful for doing experiments on large data sets, such as training sequence analysis programs.

Each record contains several important pieces. See the appendix for a sample record. Each has a unique ACCESSION number that makes searching for it simple using Entrez. A DEFINITION describes further details about the record in English. These record fields are followed by a description of the exact organism, references for who submitted the data, and other details, including publications in which it appeared. Then comes a FEATURES section, which provides details about the sequence to follow. This often includes the location of the coding sequences (CDS) for the genes within the given sequence, as well as other useful details such as the name of the protein coded, the start codon location, the location of exons within the sequence, and much more. As noted above, the complete feature definition is given in files located on the ftp site. A BASE COUNT, contaning the distribution of nucleotides in the sequence, appears just before the ORIGIN line. The line in the record following ORIGIN begins the listing of the given sequence. Each line contains 60 bases broken into 6 groups of 10 nucleotides. These lines are numbered, to facilitate fast searching for a specific position in the sequence. The sequence is followed by // which marks the end of the record.

4 Downloading Files

This section describes in slightly greater detail how to download GenBank files directly to your computer.

4.1 Web Browser

If you have a web browser, simply type `ftp://ncbi.nlm.nih.gov/genbank` into your browser and it will connect for you. If you then click on a file with the right mouse button, and choose *Save Link As …* , you will be asked where to save the file. Choose a location, click save, and the file will be downloaded. The files are plain text files that have been compressed with the `compress` utility. You will have to transfer the files to a UNIX computer to decompress them. Some web browsers will automatically decompress them in the download process, or by left clicking on the file link.

4.2 Direct Ftp

To directly ftp the files you must first login to a UNIX machine. From the command prompt type `ftp ncbi.nlm.nih.gov`. It will prompt you for a username and password and you should type `anonymous` and then use your electronic address for the password. Once logged in, type `binary` and press enter. Next, type `cd genbank` and press enter. You are now in the genbank directory which includes the GenBank release notes and sequence files. Type `ls` to see what is in the directory. Type `get filename` to download a particular file. The `README` file and the `gbrel.txt` file contain descriptions of the other files located in this directory and subdirectories. Type `bye` when your are done downloading files.

4.3 Uncompressing

Once you have successfully downloaded a file and transfered it to a UNIX machine, type `uncompress filename` to uncompress the file. For example, if you downloaded `gbest1.seq.Z`, then type `uncompress gbest1.seq.Z`. This will create a file `gbest1.seq` which is a plain text file containing the GenBank records.

5 Submission

The NCBI web site also allows for submission of new sequence records. These can be submitted either directly using online forms, or by using a program provided by the NCBI. The recommended submission method is the online method called *Bankit*, available at `http://www.ncbi.nlm.nih.gov/BankIt/`. Further details are provided on this page. The other methods, including a standalone program called *Sequin*, are further described on the GenBank home page, `http://www.ncbi.nlm.nih.gov/Genbank/`.

6 Appendix: a Sample Record

```
LOCUS       AF192387     1790 bp    mRNA            MAM       04-MAR-2000
DEFINITION  Felis catus feline leukemia virus subgroup C receptor (FLVCR1)
            mRNA, complete cds.
ACCESSION   AF192387
VERSION     AF192387.1  GI:7157951
KEYWORDS    .
SOURCE      cat.
  ORGANISM  Felis catus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Teleostomi;
            Euteleostomi; Mammalia; Eutheria; Carnivora; Fissipedia; Felidae;
            Felis.
REFERENCE   1  (bases 1 to 1790)
  AUTHORS   Quigley,J.G., Burns,C.C., Anderson,M.M., Lynch,E.D., Sabo,K.M.,
            Overbaugh,J. and Abkowitz,J.L.
  TITLE     Cloning of the cellular receptor for feline leukemia virus subgroup
            C (FeLV-C), a retrovirus that induces red cell aplasia
  JOURNAL   Blood 95 (3), 1093-1099 (2000)
  MEDLINE   20115198
   PUBMED   10648427
REFERENCE   2  (bases 1 to 1790)
  AUTHORS   Quigley,J.G., Lynch,E.D., Overbaugh,J. and Abkowitz,J.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-OCT-1999) Department of Medicine, University of
            Washington, 1959 Pacific Avenue NE, Seattle, WA 98195-7710, USA
FEATURES             Location/Qualifiers
     source          1..1790
                     /organism="Felis catus"
                     /db_xref="taxon:9685"
                     /cell_line="3201B"
                     /cell_type="CD4-positive T lymphoid"
     gene            1..1790
                     /gene="FLVCR1"
     CDS             42..1724
                     /gene="FLVCR1"
                     /note="feFLVCR; member of the major facilitator
                     superfamily of transporters"
                     /codon_start=1
                     /product="feline leukemia virus subgroup C receptor"
                     /protein_id="AAF37351.1"
                     /db_xref="GI:7157952"
                     /translation="MVKLNDEEGAAMAPGHQPTNGYLLVPGGEPPGKVSAELQNGPKA
                     VCLTLNGVSRDSLAAAAEALCRPQTPLAPEEETQTRLLPTGPGEETPGTEGSPAPQTA
                     LSARRFVVLLIFSLYSLVNAFQWIQYSVISNVFEGFYGVSSLHIDWLSMVYMLAYVPL
                     IFPATWLLDTRGLRLTALLGSGLNCLGAWVKCASVQQHLFWVTMLGQCLCSVAQVFIL
                     GLPSRIASVWFGPKEVSTACATAVLGNQLGAAIGFLLPPVLVPNTQNNTDLLACNIST
                     MFYGTSSVATFLCFLTIIAFKEKPQYPPSQAQAALQNSPPAKYSYKKSIRNLFRNVPF
                     VLLLITYGIITGAFYSVSTLLNQMILTYYKGEEVSAGKIGLTLVVAGMVGSILCGFWL
                     DYTKIYKQTTLIVYILSFLGMVIFTFTLDLGYGIVVFVTGGVLGFFMTGYLPLGFEFA
                     VEITYPESEGTSSGLLNAAAQIFGILFTLAQGKLTTDYSPKAGNIFLCVWLFLGIILT
                     ALIKSDLRRHNINIGIANGDIKAVPVEDTVEDSPTDKESKTIVMSKQSESAI"
BASE COUNT      426 a    450 c    449 g    465 t
ORIGIN
        1 atccagtgtg ctggaaaggc ggcgcagggg accccaggga catggtgaag ctgaacgatg
       61 aggagggggc agcgatggca cccgggcacc agcccacgaa tggatacctc ctggtgccgg
      121 gaggcgagcc ccccggaaag gtgagcgccg agctgcagaa cgggcccaaa gccgtctgcc
      181 tgaccctgaa tggagtgtct cgggacagcc tcgctgccgc ggcggaagcc ctgtgcaggc
```

```
 241 cgcagactcc gttggctcca gaggaggaga cccagactcg gctgctgccc acgggccccg
 301 gggaagagac cccggggacc gagggctccc cggctcccca gaccgcgctg tctgcgcggc
 361 ggtttgtggt gctcctgatc ttcagcctgt actcgctggt gaacgccttt cagtggatcc
 421 agtacagtgt catcagcaac gtcttcgagg gcttctacgg cgtctcctcc ctgcacattg
 481 actggctgtc catggtgtac atgctggcct acgtgcccct catcttcccg gccacgtggc
 541 tgctggacac cagaggccta cggctcaccg ccttgctggg ctcaggcctc aactgcctgg
 601 gcgcctgggt caagtgcgcc agcgtgcagc agcatctctt ctgggtcacc atgctgggcc
 661 agtgcctctg ctccgtggcc caggtgttca ttctcggctt gccctcccgc atcgcctcag
 721 tgtggtttgg gcccaaggag gtatccacgg cttgtgccac cgccgtgcta ggcaatcagc
 781 ttggagctgc cattggcttt ttgctgccac cggttttagt gcccaacacg cagaataaca
 841 cagatcttct ggcctgtaat atcagcacca tgttttatgg aacatcatct gttgccacgt
 901 ttttatgttt tttaacaata attgcattca aagaaaaacc tcagtatcca ccaagtcagg
 961 ctcaagcagc tcttcaaaac agcccccctg ctaagtactc ctataagaaa tcaataagga
1021 acctatttag aaacgttccc tttgtccttc tattgatcac ttatggtatc ataactggag
1081 cattttattc agtttcaaca ttattgaatc aaatgatatt gacatattac aagggagaag
1141 aagtgagtgc tggaaaagatt gggctaacat tggtagtggc tggaatggtg ggctctattc
1201 tttgtggctt ttggcttgat tataccaaaa tatacaaaca gactactctg attgtttaca
1261 ttttatcttt tcttggaatg gttatattta ctttcacatt ggaccttgga tacggtatcg
1321 ttgtgtttgt tactggaggg gtgcttggtt tcttcatgac tggttacctc ccattgggtt
1381 ttgaatttgc cgttgaaatc acttaccctg aatctgaagg cacctcatct ggtcttctta
1441 atgctgctgc acagatattt ggaattctgt tcacattggc tcaaggaaaa ctcacaacag
1501 actatagtcc taaagcagga aacattttcc tctgtgtttg gttgtttcta ggcatcattt
1561 taacagcatt aatcaagtct gatctcagaa gacacaatat aaatatagga attgcaaatg
1621 gtgatattaa agctgtacca gttgaagata cagttgaaga tagtcccaca gataaagaat
1681 caaaaactat tgtgatgtcc aagcagtcag aatcggcaat ttgaggtgaa aaaaaaactt
1741 tccagcacag tggtcgacga taaaataaaa gaacctcgag agatctaatt
//
```