**Data format in Bioinformatics**

Sequence data and associated data formats are important to bioinformaticists. This tutorial will guide you through the most frequently used data formats in bioinformatics analysis.

(1) **FASTA format**

In bioinformatics, **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotide or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

For example:

>gi|5524211|gb|AAD44166.1| cytochrome b

LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV

EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG

LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL

GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX

IENY

Here, the first “>” is mandatory and immediately followed by sequence name. In above example, the sequence name is “gi|5524211|gb|AAD44166.1|”. One or more spaces are followed immediately after the sequence name. The first space is used to separate the sequence name and the comments. In addition, sequence itself can occupy one or multiple lines. For more information, please read the web page (<http://en.wikipedia.org/wiki/FASTA_format>).

(2) **FASTQ format**

**FASTQ format** is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity. What is the quality score? It tells us how confidence we are in identifying the nucleotide sequences (bases). The higher the quality score, the more confident we are in our nucleotide identification.

A FASTQ file normally uses four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier (or sequence name) and an *optional* description (like a FASTA comment). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. For more information, please read the web page (<http://en.wikipedia.org/wiki/FASTQ_format>).

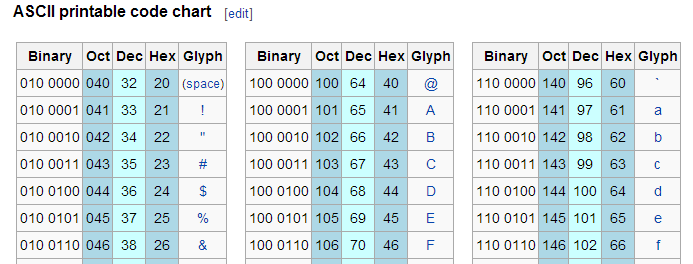
@SRR490224.1 GA2X:1:2:11987:1123/1

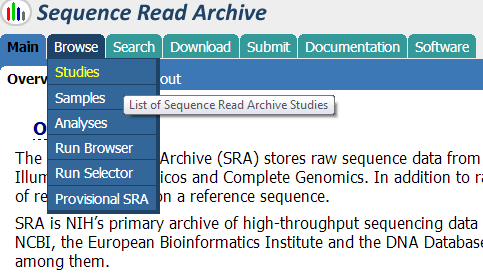
GCTCAAGGGGGGAATTCCTTCGATTCTGTTTTGCTACCCGTTGTAGCGCCGGAAGATG

+

@:::DB9>>DB@@<DBDDDC\*99@@:CBCBDD@DDBDDDDDDDDDDDDDDDDDD####

The quality value of each base is represented as single ASCII character.



(3) Two public databases that you can obtain sequences in FASTA and FASTQ files

(3.1) NCBI Trace Archive

<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>

(3.2) NCBI SRA (Sequence Read Archive)

<http://www.ncbi.nlm.nih.gov/Traces/sra/>

When you click “Browse” -> “Studies”, then you can search for “green algae” to get 112 records. Please choose the records with Accession number DRP000215 to examine and download SRA data. Then, you can download and install a software tool called NCBI SRA toolkit, which can convert the data in SRA into either FASTA or FASTQ file.

To read more about SRA toolkit, you can visit the website (<http://www.ncbi.nlm.nih.gov/books/NBK158900/>). Try this yourself!