Lecture – Sequence file formats

Sequences

Sequence formats

Why so many formats?

- ✓ There are at least a couple of dozen sequence formats in existence at the moment. Some are much more common than others.
- ✓ Formats were designed so as to be able to hold the sequence data and other information about the sequence.
- ✓ Nearly every sequence analysis package written since programs were first used to read and write sequences has invented its own format.
- ✓ Nearly every collection of sequences that dares call itself a database has stored its data in its own format.

Source: http://emboss.sourceforge.net/docs/themes/SequenceFormats.html#seq

Sequences

Sequence formats

Text files

- ✓ There are different formats to store sequences in a text file. Text files should only include Plain text.
- ✓ Graphics or any other binary information are not allowed in text files.

1. Sequences in plain files

- ✓ We store the sequence in a text file by just writing the sequence. This files include only IUPAC characters.

Example – plain format

Microsoft WORD format is not a sequence format.

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTAC
GGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGT
AGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCC
ATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
```

This kind of file is seldom used because it lacks any metadata to identify the sequence.

Note: A file in plain sequence format may only contain one sequence, while most other formats accept several sequences in one file.

Source: https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html https://www.genomatix.de/online_help/help/sequence_formats.html

Sequences

Sequence formats

2. FASTA format
(most common format)

- ✓ **The FASTA file includes a name for the sequence and, optionally, some description.**
- ✓ **The sequence should be preceded by a line that starts with the symbol >. The name will be written after that symbol.**
- ✓ **If required, several sequences can be included in the same file.**

Example

```
>sequence1_name description
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTAC
GGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGT
AGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCC
ATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
>sequence2_name description
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGA
```

Source: https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html

Sequences

Sequence formats

3. **FASTQ format**

- ✓ A sequence file in FASTQ format can contain several sequences.
- ✓ FASTQ is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.
- ✓ It is mainly used for storing the output of high-throughput sequencing instruments.

A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description.
2. the raw sequence letters.
3. a '+' character (separator), optionally followed by the same sequence identifier (and any description).
4. quality values for the sequence in Line 2.

Source: https://www.genomatix.de/online_help/help/sequence_formats.html https://sapac.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Sequences

Sequence formats

3. **FASTQ format**

Example

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description.
2. the raw sequence letters.
3. a '+' character (separator), optionally followed by the same sequence identifier (and any description).
4. quality values for the sequence in Line 2 (e.g. Phred +33 encoded, using ASCII characters).

Source: https://www.genomatix.de/online_help/help/sequence_formats.html
https://sapac.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

ASCII: American Standard Code for Information Interchange

Sequences

Sequence formats

4. EMBL format

- ✓ A sequence file in 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 ("//").

Example

```
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;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

Source: https://www.genomatix.de/online_help/help/sequence_formats.html

Sequences

Sequence formats

5. GenBank format

- ✓ A sequence file in GenBank format can contain several sequences.
- ✓ One sequence in 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 "ORIGIN" and the end of the sequence is marked by two slashes ("//").

Example

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        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 gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

Source: https://www.genomatix.de/online_help/help/sequence_formats.html

Sequences

Tutorial 6

Sequence formats

Tutorial

- ✓ Write down the five sequence formats?
- ✓ Which one is the most common format?

Source: https://www.genomatix.de/online_help/help/sequence_formats.html