Some data formats in bioinformatics

FASTA

DNA sequences are stored in various formats depending upon the database and application. The most common format is FASTA. In this format the sequence name and additional information is provided on one line beginning with the '>' character. The sequence itself is represented in following lines either in interleaved format, which means a fixed number of characters per line, or non-interleaved. Below are examples of interleaved and non-interleaved formats.

Interleaved FASTA (three sequences):

```
>human
ACCGTGATGT
AGAGACCACG
GGCCC
>mouse
CCCAGTGTGT
AACA
>cat
AGTGTGTGTT
GTGCCCG
```

Non-interleaved FASTA (of the above example):

```
>human
ACCGTGATGTAGAGACCACGGGCCC
>mouse
CCCAGTGTGTAACA
>cat
AGTGTGTGTTGTGCCCG
```

PHYLIP

A popular format that is used in phylogeny (evolutionary tree) reconstruction is PHYLIP. In this format the sequences must be aligned. The first line contains the number of sequences and their length – since sequences are aligned they will all have the same length. The following lines each contain the name of the sequence followed by one or more spaces and the sequence. In interleaved format the sequence is represented across several lines along with the name as well. See examples below

Non-interleaved PHYLIP (three sequences)

```
3 44
human     ACGTGTGTGACAGTGTGAGACCACGTGACCCCCCCCCCCGCGCG
mouse     ACCCGTGTGTGGGGTGTAGACCACG---CCCCCCCCCGT-----
cat       ACCCCGTGGG--------GACCACGTGACCCCCCCCAGT-----
```

Interleaved PHYLIP with 10 characters per block of each sequence (above example)

```
3 44
human     ACGTGTGTGA
mouse     ACCCGTGTGT
cat       ACCCCGTGGG

human     CAGTGTGAGA
mouse     GGGGTGTAGA
cat       --------GA

human     CCACGTGACC
mouse     CCACG---CC
cat       CCACGTGACC

human     CCCCCCCGCG
mouse     CCCCCCCGT-
cat       CCCCCCAGT-

human     CGCG
mouse     ----
cat       ----
```

Another example of PHYLIP interleaved and non-interleaved:

Non-interleaved:

```
4 14
human     ACGTGTGTGACAGT
mouse     ACCCGTGTGTGGGG
cat       ACCCCGTGGG----
dog       ACCCCGTGTG----
```

Interleaved with 10 characters per block of each sequence:

```
4 14
human     ACGTGTGTGA
mouse     ACCCGTGTGT
cat       ACCCCGTGGG
dog       ACCCCGTGTG

human     CAGT
mouse     GGGG
cat       ----
dog       ----
```