Exercise Sheet 4

Problem 1. Text processing: Nucleotide sequence
The fasta format is an ubiquitous format for representing nucleotide sequences. A typical fasta file starts with a single-line description, followed by lines of sequence data. The description line start with a > and is directly followed by the name of the sequence. Further information might be given after a blank space. A sequence ends if a new line with a > appears. More information on the fasta format can be found at `http://zhanglab.ccmb.med.umich.edu/FASTA/`

(a) The file `example.fasta` can be found on the SLC. It contains three DNA sequences. Read the data into R (e.g., `readLines`)

(b) Create a data.frame with 2 variables

`seq_name`**:** This variable should contain the sequences' id, i.e., everything after the > and before the first space

`sequence`**:** The sequence in itself

(c) Answer the following questions:

(a) How long is each sequence?

(b) How many A's, T's, G's and C's and is there in each sequence? Propose a simple graphic to display this information

(d) Transform the DNA sequences into RNA (change the T's in U's)