~~~
---
output: html_document
editor_options:
  chunk_output_type: console
---
# Downloading protein sequences in *R* {#download-protein-seqs-R}

**By**: Avril Coghlan.

**Adapted, edited and expanded**: Nathan Brouwer under the Creative
Commons 3.0 Attribution License [(CC BY
3.0)](https://creativecommons.org/licenses/by/3.0/).


<!-- Add gsub walk through -->

## Preliminaries

```{r}
library(compbio4all)
```



## Retrieving a UniProt protein sequence using rentrez

We can use `entrez_fetch()` to download protein sequences.

For example to retrieve the protein sequences for UniProt accessions
Q9CD83 and A0PQ23, we type in R:


```{r}
# sequence 1: Q9CD83
leprae_fasta <- rentrez::entrez_fetch(db = "protein",
                        id = "Q9CD83",
                         rettype = "fasta")
# sequence 2: OIN17619.1
ulcerans_fasta <- rentrez::entrez_fetch(db = "protein",
                         id = "OIN17619.1",
                         rettype = "fasta")
```



Display the contents of the `lepraeseq` FASTA file.
```{r}
leprae_fasta
```

<!-- When did I introduce FASTA cleaner? -->
<!-- Are these data files in compbio4all? -->
~~~

~~~
Let's clean these up to remove the header and new line characters usin the
function `fasta_cleaner()`.
```{r}
leprae_vector   <- fasta_cleaner(leprae_fasta)
ulcerans_vector <- fasta_cleaner(ulcerans_fasta)
```

Examine the output using `length()`, `class()`, and `head()`:

```{r, eval = F}
length(leprae_vector)
class(leprae_vector)
head(leprae_vector, 20)
```
~~~