---

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)

```