Proteínas capaces de cristalizar en un intervalo de pH amplio

Francisco Murphy Pérez

2022-05-23

Contents

```
# knitr::opts_chunk$set(eval = FALSE)
# switch global para evaluar funciones o no.
```

Chapter 1

Dependencias

La reproducción correcta de esta bitácora depende de varios programas instalados en un sistema operativo linux.

1.1 Sistema operativo

El sistema operativo usado es Fedora 36 (https://getfedora.org/) en su versión *workstation*.

```
uname -r # Imprime la versión del núcleo de linux.
```

```
## 5.17.8-300.fc36.x86_64
```

1.2 Minería de datos

La extracción de datos se realiza con `gemmi` (https://github.com/project-gemmi/gemmi), que ya viene dentro de la colección de programas de `ccp4` (https://www.ccp4.ac.uk/).

Para la limpieza y transformación de datos se usa:

- `awk` (https://www.gnu.org/software/gawk/)
- `bash` (https://www.gnu.org/software/bash/)
- `grep` (https://www.gnu.org/software/grep/)
- `R` (https://www.r-project.org/)
- `sed` (https://www.gnu.org/software/gawk/)
- `tidyverse` (https://www.tidyverse.org/)

La instalación de `R` conviene hacerla como la describo en el siguiente enlace https://murpholinox.github.io/2021/05/01/installRsansTexLivedeps.html. La instalación del `tidyverse` depende a su vez de la instalación de las siguientes librerías en el sistema operativo:

```
sudo dnf install openssl-devel libcurl-devel
```

> Advertencia: Si se tiene instalado chimerax-daily se tendrá un conflicto con `openssl-devel`. En ese caso se tiene que instalar `openssl1.1-devel` y tanto chimerax-daily como rstudio corren sin problemas.

Los programas restantes (`awk`, `bash`, `grep` y `sed`) vienen instalados por defecto en el sistema operativo usado.

1.3 Configuración

Además de la instalación correcta de los programas anteriores, se tienen que cargar las siguientes librerías de `R`.

```
library(dplyr)
library(ggplot2)
library(readr)
library(knitr)
library(kableExtra)
library(stringdist)
library(svglite)
library(bookdown)
library(rmarkdown)
library(renv)
```

1.3.1 Manejo de dependencias

El manejo de dependencias, de manera interna, se da automáticamente gracias a `renv`.

```
renv::consent() # Da permiso a renv.4
```

```
## * Consent to use renv has already been provided -- nothing to do.
```

```
renv::init() # Inicia renv.
```

```
# Para ver las dependencias.
cat renv.lock
```

1.4 Sesión

Imprime información de la sesión activa de `R`.

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-redhat-linux-gnu (64-bit)
```

```
## Running under: Fedora Linux 36 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.1
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices datasets  utils     methods   base
##
## other attached packages:
##  [1] renv_0.15.4      rmarkdown_2.14   bookdown_0.26    svglite_2.1.0
##  [5] stringdist_0.9.8 kableExtra_1.3.4 knitr_1.39       readr_2.1.2
##  [9] ggplot2_3.3.6    dplyr_1.0.9
##
## loaded via a namespace (and not attached):
##  [1] pillar_1.7.0      compiler_4.1.3    tools_4.1.3       digest_0.6.29
##  [5] viridisLite_0.4.0 evaluate_0.15     lifecycle_1.0.1   tibble_3.1.7
##  [9] gtable_0.3.0      pkgconfig_2.0.3   rlang_1.0.2       cli_3.3.0
## [13] rstudioapi_0.13   parallel_4.1.3    yaml_2.3.5        xfun_0.30
## [17] fastmap_1.1.0     xml2_1.3.3        httr_1.4.3        withr_2.5.0
## [21] stringr_1.4.0     systemfonts_1.0.4 hms_1.1.1         generics_0.1.2
## [25] vctrs_0.4.1       webshot_0.5.3     grid_4.1.3        tidyselect_1.1.2
## [29] glue_1.6.2        R6_2.5.1          fansi_1.0.3       tzdb_0.3.0
## [33] purrr_0.3.4       magrittr_2.0.3    scales_1.2.0      ellipsis_0.3.2
## [37] htmltools_0.5.2   rvest_1.0.2       colorspace_2.0-3  utf8_1.2.2
## [41] stringi_1.7.6     munsell_0.5.0     crayon_1.5.1
```

1.5 Contacto

Me puedes contactar por correo electrónico en gmail o ibt.

Chapter 2

Hello bookdown

All chapters start with a first-level heading followed by your chapter title, like the line above. There should be only one first-level heading (`#`) per .Rmd file.

2.1 A section

All chapter sections start with a second-level (`##`) or higher heading followed by your section title, like the sections above and below here. You can have as many as you want within a chapter.

An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a `{.unnumbered}` or the shorter `{-}` at the end of the heading, like in this section.

Chapter 3

Introducción

El proyecto de doctorado consiste en analizar el efecto que tiene el pH en el daño por radiación en cristales de proteína. Para ello se tiene que cristalizar algunas proteínas a diferente pH.

> Todo lo demás tiene que ser idéntico: condición de cristalización (ambiente químico), grupo espacial (conformación espacial), parámetros de colecta de datos (dosis de radiación absorbida).

3.1 Objetivos

1. Obtener una lista de proteínas que cumplan los requisitos *adecuados* para llevar a cabo dicho proyecto.
2. Cristalizar las proteínas seleccionadas a diferentes niveles de pH.
3. Difractar los cristales obtenidos.
4. Realizar un análisis comparativo del daño por radiación a diferente pH.

3.2 PDB

El PDB (https://www.rcsb.org/) es un repositorio de coordenadas de macromoléculas biológicas. La mayoría de las coordenadas se obtienen por difracción de rayos-X de cristales macromoleculares. En el cabezal de cada archivo de coordenadas, se tiene la información necesaria para reproducir el experimento de cristalización de la macromolécula de interés. En otras palabras, la información necesaria para el primer objetivo se encuentra en los cabezales de los archivos.

Chapter 4

Extracción de datos

4.1 Formato

El PDB ofrece descargar sus archivos en tres formatos diferentes: `.xml`, `.pdb` y `.mmcif`. El segundo es el más fácil de leer y manipular; sin embargo, se decidió usar el tercer formato debido al siguiente párrafo:

> Many of the errors have been fixed in the equivalent mmCIF files. Hence, if you are interested in the header information, it is a good idea to extract information from mmCIF files…
>
> De https://biopython.readthedocs.io/en/latest/chapter_pdb.html.

El formato `.mmcif` se detalla en http://mmcif.wwpdb.org/. Existe una correspondencia entre las etiquetas del `.pdb` con las etiquetas del `.mmcif`.

4.2 Descarga

Para descargar todo las estructuras del PDB en formato `.mmcif`, se usa el siguiente comando:

```
cd /run/media/murphy/lolita/doctorado
rsync -avPz --delete data.pdbjbk1.pdbj.org::ftp_data/structures/divided/mmCIF/ ./mmCIF
# Tarda entre 5 y 6 horas con una buena conexión de internet.
```

> Instrucciones de https://www.wwpdb.org/ftp/pdb-ftp-sites.

4.3 Organización de archivos

Los archivos están organizados en diferentes subdirectorios, cuyo nombre está formado por el segundo y el tercer carácter del nombre del mismo archivo. Por

ejemplo `1abc.mmcif` estará en el subdirectorio `ab/`. Se realiza una copia de los archivos en un solo directorio en el disco duro con dos objetivos en mente: tener un respaldo y manipular de una manera más sencilla los archivos.

```
cd /run/media/murphy/lolita/doctorado
mkdir mmCIF_backup
cd mmCIF/
time find . -name '*.gz' -exec cp \{\} /run/media/murphy/lolita/doctorado/mmCIF_backup/
# Esto mucho no sé por qué!
```

4.3.1 Separa entradas por método experimental

De los archivos depositados en el PDB, obtenemos aquellas estructuras determinadas únicamente por difracción de rayos-X de cristales. Este se puede considerar como el primer filtro. Además ayuda a eliminar confusiones posteriores.

> El problema es que `gemmi` extrae etiquetas de manera excelente, pero no conoce contextos. Esto puede resultar, dependiendo de las etiquetas, en datos incompletos.

```
cd /run/media/murphy/lolita/doctorado/
mkdir xray
time gemmi grep _exptl.method mmCIF_backup/ > xray/method.list
# Esto tarda 155 minutos
# Se confirma con:
# wc -l method.dat
# 190846
# La diferencia con el total de entradas en el PDB (190639), es por los pdbs obtenidos
# La siguiente línea nos da donde se da esta diferencia.
# awk -F ":" '{print $1}' method.dat | uniq -c | awk '{ if ($1!="1") print $0}' | wc -
# 205
# Lo cual se confirma en la búsqueda avanzada del PDB escogiendo como método experimen
cd /run/media/murphy/lolita/doctorado/xray/
grep X-RAY method.list | awk -F : '{print $1}' | tr '[:upper:]' '[:lower:]' > pdbs_by_x
sed 's/$/.cif.gz/'g pdbs_by_xray.list > list_pdbs_by_xray
# Es interesante comparar el total de entradas en el PDB con aquellas obtenidas por di
# wc -l pdbs_by_xray.dat
# 165662
mkdir entries
time cat list_pdbs_by_xray | while read line;
do cp /run/media/murphy/lolita/doctorado/mmCIF_backup/$line entries/; done
# Esto tarda 145 minutos.
```

4.4 Extracción de datos

```
# Usar un delimitador que no aparece en los archivos.
cd /run/media/murphy/lolita/doctorado/xray/
time gemmi grep --delimiter='¿' _entity_poly.entity_id -a _entity_poly.type -a _struct_ref.pdbx_d
# Esto tarda 52 minutos.
# wc -l information_from_xrays
# 255251
# La diferencia con el total de entradas en el PDB, es porque varios archivos contienen más de un
# Por ejemplo:
# 10MH¿1¿polydeoxyribonucleotide¿P05102¿DNA (5'-D(P*CP*CP*AP*TP*GP*(5CM)P*GP*CP*TP*GP*AP*C)-3')¿¿
# 10MH¿2¿polydeoxyribonucleotide¿P05102¿DNA (5'-D(P*CP*CP*AP*TP*GP*(5CM)P*GP*CP*TP*GP*AP*C)-3')¿¿
# 10MH¿3¿polypeptide(L)¿P05102¿DNA (5'-D(P*CP*CP*AP*TP*GP*(5CM)P*GP*CP*TP*GP*AP*C)-3')¿¿6.5¿10% (
```

4.5 Verifica la integridad de los datos y obtiene proteínas más representadas en el PDB

Importa los datos extraídos a `R` y realiza algunas gráficas interesantes.

```
library(readr)
da<-read_delim("/home/murphy/doctorado/info.txt", delim = "¿", escape_double = FALSE, col_names =
```

```
## Rows: 255251 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "¿"
## chr (8): X1, X3, X4, X5, X6, X8, X11, X12
## dbl (4): X2, X7, X9, X10
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df0 <- da %>%
  rename(
    pdb = X1,
    nde = X2,
    tde = X3,
    ide = X4,
    nom = X5,
    tec = X6,
    peh = X7,
    con = X8,
    rs1 = X9,
    rs2 = 10,
    gpo = X11,
    doi = X12
```