Pre-processing and plotting data

Mike Blazanin

Contents

Where are we so far? .......................... 1

Pre-processing ................................ 2
  Pre-processing: excluding data ......... 2
  Pre-processing: converting dates & times into numeric .... 2

Plotting your data ............................ 4

What’s next? .................................. 5

Where are we so far?

1. Introduction: vignette("gcplyr")
2. Importing and transforming data: vignette("import_transform")
3. Incorporating design information: vignette("incorporate_designs")
4. Pre-processing and plotting your data: vignette("preprocess_plot")
5. Processing your data: vignette("process")
6. Analyzing your data: vignette("analyze")
7. Dealing with noise: vignette("noise")
8. Statistics, merging other data, and other resources: vignette("conclusion")

So far, we’ve imported and transformed our measures, then combined them with our design information. Now we’re going to do some final pre-processing steps and show how to easily plot our data with ggplot.

If you haven’t already, load the necessary packages.

```r
library(gcplyr)
library(dplyr)
library(ggplot2)
library(lubridate)
```

#> Attaching package: 'lubridate'
#> The following objects are masked from 'package:base':
#>
#> date, intersect, setdiff, union
Pre-processing

Now that we have our data and designs merged, we’re almost ready to start processing and analyzing them. However, first we need to carry out any necessary pre-processing steps, like excluding wells that were contaminated or empty, and converting time formats to numeric.

Pre-processing: excluding data

In some cases, we want to remove some of the wells from our growth curves data before we carry on with downstream analyses. For instance, they may have been left empty, contained negative controls, or were contaminated. We can use dplyr’s filter function to remove those wells that meet criteria we want to exclude.

For instance, let’s imagine that we realized that we put the wrong media into Well B1, and that strain 13 was contaminated. To exclude them from our analyses, we can simply:

```r
example_data_and_designs_filtered <- filter(ex_dat_mrg,
                                            Well != "B1", Bacteria_strain != "Strain 13")
head(example_data_and_designs_filtered)
```

<table>
<thead>
<tr>
<th>Time</th>
<th>Well</th>
<th>Measurements</th>
<th>Bacteria_strain</th>
<th>Phage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>0.002</td>
<td>Strain 1</td>
<td>No Phage</td>
</tr>
<tr>
<td>2</td>
<td>D1</td>
<td>0.002</td>
<td>Strain 19</td>
<td>No Phage</td>
</tr>
<tr>
<td>3</td>
<td>E1</td>
<td>0.002</td>
<td>Strain 25</td>
<td>No Phage</td>
</tr>
<tr>
<td>4</td>
<td>F1</td>
<td>0.002</td>
<td>Strain 31</td>
<td>No Phage</td>
</tr>
<tr>
<td>5</td>
<td>G1</td>
<td>0.002</td>
<td>Strain 37</td>
<td>No Phage</td>
</tr>
<tr>
<td>6</td>
<td>H1</td>
<td>0.002</td>
<td>Strain 43</td>
<td>No Phage</td>
</tr>
</tbody>
</table>

Pre-processing: converting dates & times into numeric

Growth curve data produced by a plate reader often encodes the timestamp information as a string (e.g. “2:45:11” for 2 hours, 45 minutes, and 11 seconds), while downstream analyses need timestamp
information as a numeric (e.g. number of seconds elapsed). Luckily, others have written great packages that make it easy to convert from common date-time text formats into plain numeric formats. Here, we’ll see how to use lubridate to do so:

First we have to create a data frame with time saved as it might be by a plate reader.

```r
#Don’t worry how this code works, it’s just creating an example file
# in the same format that a plate reader would
ex_dat_mrg$Time <-
  paste(ex_dat_mrg$Time %% 3600,
        formatC((ex_dat_mrg$Time %% 60) %% 60,
            width = 2, flag = 0),
        formatC((ex_dat_mrg$Time %% 3600) %% 60,
            width = 2, flag = 0),
        sep = ";")
head(ex_dat_mrg)
#>  Time Well Measurements Bacteria_strain Phage
#> 1 0:00:00 A1 0.002 Strain 1 No Phage
#> 2 0:00:00 B1 0.002 Strain 7 No Phage
#> 3 0:00:00 C1 0.002 Strain 13 No Phage
#> 4 0:00:00 D1 0.002 Strain 19 No Phage
#> 5 0:00:00 E1 0.002 Strain 25 No Phage
#> 6 0:00:00 F1 0.002 Strain 31 No Phage
```

We can see that our Time aren’t written in an easy numeric. Instead, they’re in a format that’s easy for a human to understand (but unfortunately not very usable for analysis).

Let’s use lubridate to convert this text into a usable format. lubridate has a whole family of functions that can parse text with hour, minute, and/or second components. You can use hms if your text contains hour, minute, and second information, hm if it only contains hour and minute information, and ms if it only contains minute and second information.

Once hms has parsed the text, we’ll use time_length to convert the output of hms into a pure numeric value. By default, time_length returns in units of seconds, but you can change that by changing the unit argument to time_length.

```r
# We have previously loaded lubridate, but if you haven’t already then
# make sure to add the line:
# library(lubridate)

ex_dat_mrg$Time <- time_length(hms(ex_dat_mrg$Time), unit = "hour")
head(ex_dat_mrg)
#>  Time Well Measurements Bacteria_strain Phage
#> 1 0 A1 0.002 Strain 1 No Phage
#> 2 0 B1 0.002 Strain 7 No Phage
#> 3 0 C1 0.002 Strain 13 No Phage
#> 4 0 D1 0.002 Strain 19 No Phage
#> 5 0 E1 0.002 Strain 25 No Phage
#> 6 0 F1 0.002 Strain 31 No Phage
```

And now we can see that we’ve gotten nice numeric Time values! So we can proceed with the next steps of the analysis.
Plotting your data

Once your data has been merged and times have been converted to numeric, we can easily plot our data using the ggplot2 package. That’s because ggplot2 was specifically built on the assumption that data would be tidy-shaped, which ours is! We won’t go into depth on how to use ggplot here, but there are three main commands to the plot below:

- **ggplot** - the ggplot function is where you specify the data.frame you would like to use and the aesthetics of the plot (the x and y axes you would like)
- **geom_line** - tells ggplot how we would like to plot the data, in this case with a line (another common geom for time-series data is geom_point)
- **facet_wrap** - tells ggplot to plot each Well in a separate facet

We’ll be using this format to plot our data throughout the remainder of this vignette

```r
# We have previously loaded ggplot2, but if you haven't already then
# make sure to add the line:
#   library(ggplot2)

# First, we'll reorder the Well levels so they plot in the correct order
ex_dat_mrg$Well <-
  factor(ex_dat_mrg$Well,
         levels = paste(rep(LETTERS[1:8], each = 12), 1:12, sep = ""))

ggplot(data = ex_dat_mrg, aes(x = Time, y = Measurements)) +
  geom_line() +
  facet_wrap(~Well, nrow = 8, ncol = 12)
```
Generally speaking, **from here on you should plot your data frequently**, and in every way you can think of! **After every processing and analysis step, visualize both the input data and output data** to understand what the processing and analysis steps are doing and whether they are the right choices for your particular data (this vignette will be doing that too!)

**What’s next?**

Now that you’ve pre-processed and visualized your data, it’s time to process (in most cases) and analyze (pretty much always) it!

1. Introduction: `vignette("gcplyr")`
2. Importing and transforming data: `vignette("import_transform")`
3. Incorporating design information: `vignette("incorporate_designs")`
4. Pre-processing and plotting your data: `vignette("preprocess_plot")`
5. Processing your data: `vignette("process")`
6. Analyzing your data: `vignette("analyze")`
7. Dealing with noise: `vignette("noise")`
8. Statistics, merging other data, and other resources: `vignette("conclusion")`