

ASCA Exercises 2019

Biosystems Data Analysis

Age Smilde, Johan Westerhuis, Gooitzen Zwanenburg

**Introduction**

Plants in the Cabbage family (Brassicaceae) make very specific defense compounds when under attack by herbivores. These compounds are called glucosinolates: about 120 different species exist, varying only in the group R (Figure 1). To simulate a herbivore attack and elicit the plant to make glucosinolates, the plant hormone jasmonic acid (JA) was applied to either the roots (Root induced) or to the leaves (Shoot induced) of cabbage plants (B. oleracea). The glucosinolate levels for 11 different glucosinolates were measured at 1, 3, 7 and 14 days after treatment. This measurement was destructive so that for each time-point different plants were analyzed. Each measurement was repeated 5 times. Besides the two treatments there was also a control group that received no JA treatment. We would like to know how the different treatments affect the glucosinolate composition and whether the glucosinolate composition varies with time.

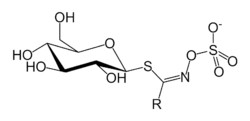


Figure 1 Glucosinolate

Table 1 shows the glucosinolates that were measured and the corresponding number of the variable in the data matrix.

|  |  |
| --- | --- |
| Variable number | Glucosinolate |
| 1 | PRO |
| 2 | RAPH |
| 3 | ALY |
| 4 | GNL |
| 5 | GNA |
| 6 | 4OH |
| 7 | GBN |
| 8 | GBC |
| 9 | 4MeOH |
| 10 | NAS |
| 11 | NEO |
|  |  |

Table 1*Variables and abbreviations of the glucosinolates that were measured in the experiment.*

**Exploring the data**

***Some preliminaries***

To explore the data, we use a data exploration tool written in R. To run the tool, you need to have the latest version of the R programming language and the R programming environment RStudio installed. Furthermore, you need to install the packages *ggrepel, shiny, shinydashboard, tidyverse and plyr*. The data exploration tool and the data we want to study can be downloaded from Canvas. Download the file *Datatool.zip* from Canvas and unpack the file in a folder of your choice. In the folder where you unpacked the zip-file you’ll find five files with a .R extension a file called *Data.csv* and a file *F.csv*.

*Data.csv* contains the measured glucosinolate levels, the first row of the file gives the names of the glucosinolates. Each row corresponds to the measurement of the glucosinolate levels of a root, shoot or control sample. The file *F.csv* tells you to which measurement the rows correspond. The first column indicates the time point at which the measurement was taken, the second column gives the method: root induced, shoot induced or control. In this way we can see that, for example, the sixth row of the data file corresponds to root induced JA, measured at Day 1.

**Exercise 1**

We begin by starting the data tool. Go to the folder where you unpacked *Datatool.zip* and open the file *DataTool.R* in RStudio (see Figure 2). Start the data tool by clicking the *Run App* button in the top right corner of the top left window. The data tool will now open in a web browser or in a new RStudio window, depending your setting. You can change where the data tool opens by opening a menu by clicking the little black triangle next to the *Run App* text.

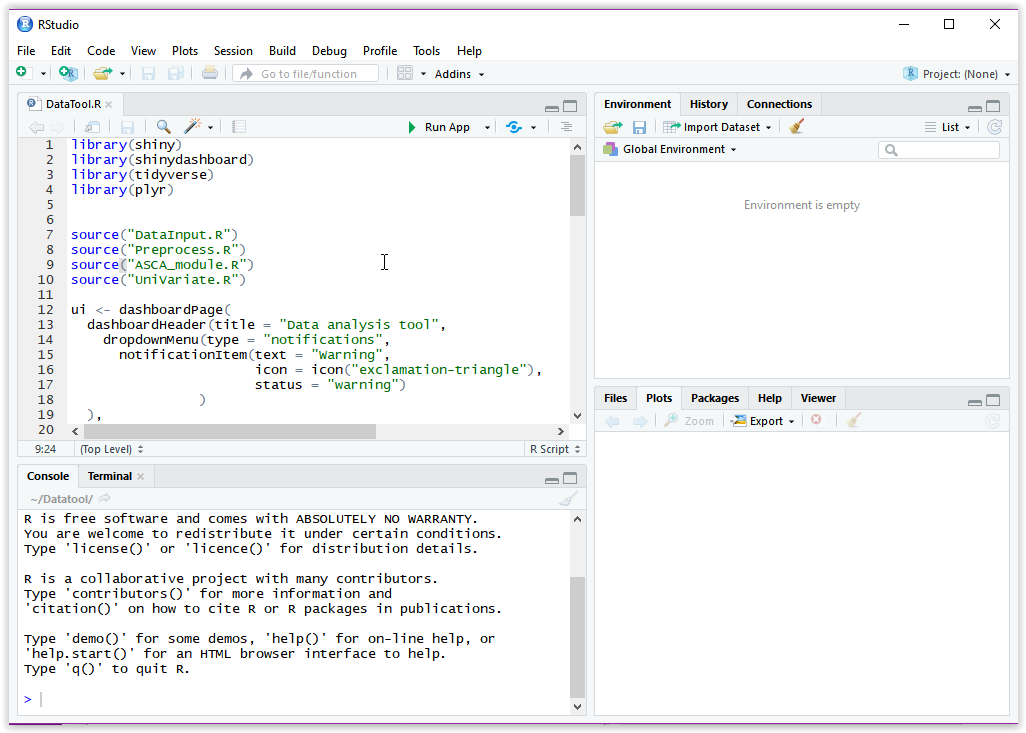


Figure 2 RStudio after opening the file DataTools.R

Figure 3 shows the data tool opened in a RStudio window.

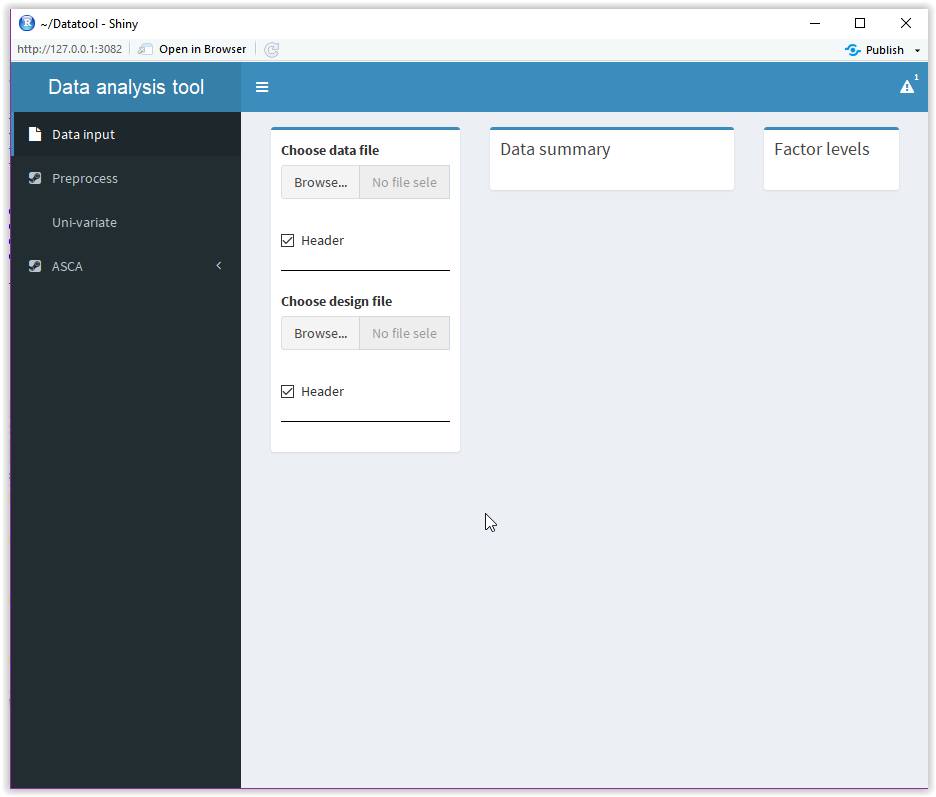


Figure 3 The data tool opened in a new RStudio window.

Use the **Choose data file** menu to read in the file *Data.csv* and the **Choose design file** menu to read in the file *F.csv*. Both files have a `header’ (the first row with the names of the glucosinolates and the experimental factors), so the **Header** box is ticked for both files. If all goes well, the data and the design are summarized in the two tables to the right.

* How many glucosinolates are measured?
* What are the levels of the factor *Time?*
* What are the levels of the factor *Method?*
* How many samples were taken for each *Time/Method* combination?

To continue, we need to click the **Preprocess**menu in the left column. With this menu we can scale the data or, if the data is unbalanced, balance the data. Because the glucosinolate data are balanced, we will not use the **Balancing** menu. The **Scaling** menu, allows you to apply different scaling methods to the data. If the **Center** box is ticked, the data will be centered (the default).

With the **Plot** menu, you can choose to plot the scores or the loadings, or a biplot in which the scores and the loadings are combined in a single plot. If you select *Data*, the measurements for the individual variables are displayed. Select *Scores* andbelow the menus, you see a principal component (PCA)-plot of the whole data set. Such a plot gives you a first idea of outliers and grouping in the data. A loadings plot can give you an idea which variables are ‘important’. The biplot makes the connection between the score values and the loadings. With the **Color by** menu you can identify which points belong to the different factor levels.

* Try different scaling methods and see how they affect the PCA plot of the data.

**Exercise 2**

To look at the individual variables, we use the **Univariate** menu. For each variable, you can put the different factor levels on the x-axis; the measured values are represented by a box plot. In this way, you can see how the individual variables behave. By selecting multiple items from the **Variables** menu, you can put different glucosinolates side by side. You can remove a variable from the list by selecting it and pressing the *delete*-button.

* Discuss the problematic issues, specific for metabolomics data that you observe for this data.

Next we will look in more detail at the concentrations measured for each of the glucosinolates. The measured values (small dots) and averages (large dots) of the five repeated measurements are also shown in the box plots. From the **X-axis** menu, select *Time*. The levels for the factor *Time* are now plotted on the x-axis. From the **Color by** menu, select *Treatment*. The root induced, shoot induced and control groups are now colored differently and their averages connected by a line.

* From the box plot you see that some variables clearly show the effect of the treatment while others seem unaffected. Which variables show a clear treatment effect?

After seeing the plots for the individual variables we have an idea what the data look like. However, the plots of the individual variables make it difficult to see which variables contribute most to the variation in the data and to spot outliers. Therefore, we return to principal component analysis to take a look at the overall variation in the data. We noticed from the box plots that the variables are all measured in the same units and their values are all of the same order of magnitude.

**Exercise 3**

* Why is it important for PCA that the measured values of all variables are approximately of the same magnitude?
* Why is centering a good idea before doing PCA?

Go back to the **Preprocess** menu.

* Do PCA on the data without centering. How can you see from the score plot that the data are not centered?

Now check the **Center** box to center the data.

* Which variables contribute most to the variation? If this surprises you, have another look at the box plots.

Considering again the box plots for the individual variables, you can see that the assumption of equal variance within a group is violated. To make the variances within the groups more equal one can use a logarithmic or square root transformation of the data. We select the square root transformation.

**Exercise 4**

Select *Square root* from the **Transformation** menu. This operation takes the square root of the measured data. If the measurement happens to be negative, the square root of its absolute value is taken and the resulting value multiplied by -1.

* Compare the variance of the individual variables with and without taking the square root of the data.
* Also compare the PCA score and loading plots (biplot) for data with and without taking the square root for both centered and non-centered data.
* What is your conclusion?

From now on, we are going to work with the square root transformed data and start using ASCA on this data.

**Exercise 5**

* Sketch the experimental design in the same way as the top part in Figure 1 of the ASCA paper.
* Write down the decomposition of the data matrix in effect and interaction matrices. How many treatment levels does each of the matrices have?
* In the PCA plot there were a few points that could be considered outliers. Outliers have less effect on ASCA than on regular PCA, why?

To do ASCA with the Datatool we click the **ASCA** menu in the sidebar. The first thing we need to do is build a model. The model tells us which factors, interactions or combinations of the two to include. If, for example we want to include only the two factors, we type *1,2* in the **Include factors** box. The model is now:

X = A + B + R

with X is the data, A is the first factor (Time), B is the second factor (Treatment) and R is the residual matrix.

**Exercise 6**

Enter the above model in the Datatool.

* Which factor explains most of the variation?
* How much of the variation is in the residuals?
* Notice that these values hold for the data as how it was preprocessed in the preprocessing menu (thus square root transformation and centered). Check how the values change when the data are not centered after square root transformation.

You can view the scores and loadings plots for the different model components under the **Plots** menu in the sidebar. As no interaction have been calculated yet, it is only possible to see the scores of the Factors and of the Residuals. The score plots generated have more points than you probably expected to see. There are so many points because not only the scores of the level averages are plotted, but also the projection of the data points onto the space spanned by the two principal components that are plotted. In this way we get a better idea of the distribution of the data points. If you uncheck the Projections box, then the projections of the data points disappear. Try out the different plot options for Scores, Levels, Loadings and Biplots.

**Exercise 7**

Do ASCA on the square root transformed and centered data without interaction.

* How much variation is explained by each of the factors?
* Seeing the score plots, what would you say about the effect of the two treatments (time and application)?
* Compare the results you find with ASCA with those you found by observing the box plots. Especially pay attention to the treatment effect on the different glucosinolates.

**Exercise 8**

We did not include interactions between the two experimental factors yet.

* Before doing ASCA with interactions, do you expect a strong interaction effect?

To check your opinion, you need to include the interaction between the two factors into the model. You do this by entering *1:2* into the **Include interactions** box.

* How much of the variation is explained by the interactions?
* Answer the question we posed at the start of the exercises: “We would like to know how the different treatments affect the glucosinolate composition and whether the glucosinolate composition varies with time”.

If, for example, the dose and dose-time interaction matrices are combined: **X***β* + **X**(*αβ*) we can look at the effect of the dose application at different times. In the Datatool we can do this in the **Combine terms** menu under the **Build Model** tab in the side bar. We enter in the **Combine terms** box: *2+1:2* to indicate that we want to add factor *2* and the interaction between *1* and *2.* Notice that when we do this, that the model changes and the interactions and factor *2* are no longer part of the model. This makes sense, because we can use each term in the model only once!

Go through the different plot options of the Combination and try to interpret the different plots (e.g. look at a level plot of PC1 with Factor 1 on x-axis and colored by Factor 2).

**Exercise 9**

* This is a point for discussion. What are we looking at when we combine two effects as we did above?
* Make plots of only the treatment effect and only the interaction effect in a similar way as presented above. Discuss the difference with the plot above.