Plant PhysioSpace Tutorial

Ali H. Esfahani

May 2016

1 Introduction

Plant PhysioSpace is an application for analyzing and comparing gene expression data acquired from plants. The initial version of this application is implemented based on Human PhysioSpace [1] with efforts of David Nevarez in his master thesis done in [2]

This tutorial teaches you everything you need to know for using the web app version of Plant PhysioSpace accessible under `http://dev.plabipd.de/physiospace`.

2 Basic look and Home screen

When you open `http://dev.plabipd.de/physiospace` for the first time, you see figure 1. If you already have an account, you can log in using the link in the light blue bar. If not, as explained in the grey box in the middle of figure 1, you can either go to "Submit List of Genes" tab, or use "Submit csv".
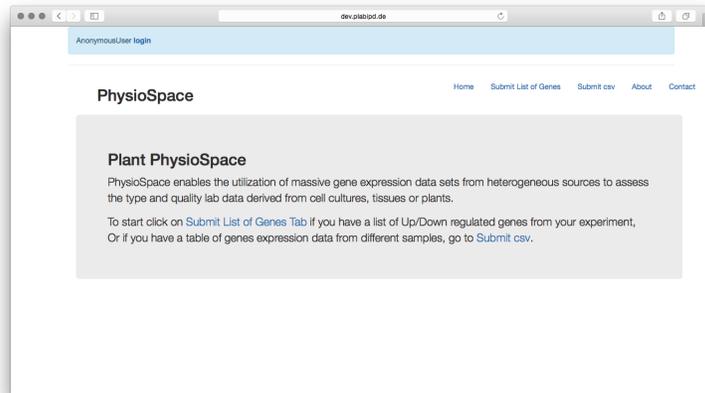
Figure 1: Home Screen

2.1 "Submit List of Genes" section

This section is a more simplified and easier to use version of the program. In this section, you can compare your samples to a compendium generated using mainly "The AtGenExpress global stress expression data set" [3] among others.

Opening the "Submit List of Genes" tab, you should see figure 2. Using this
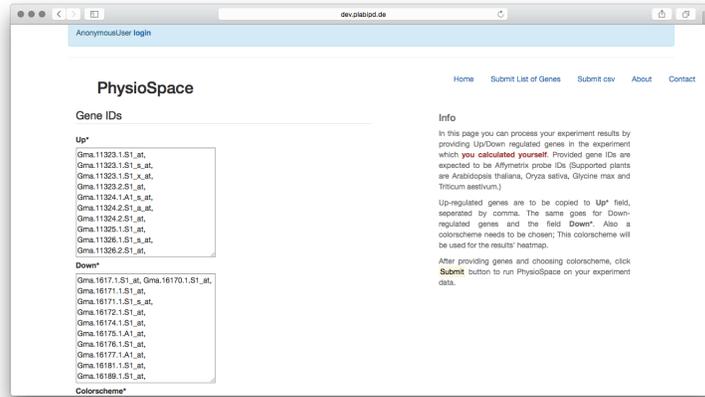
Figure 2: Gene list tab

tab you need to take these two steps:

1. This section processes each sample separately. For each sample, you have to calculate Up and Down regulated genes and find their corresponding **the Affymetrix gene IDs** and paste those IDs in `Up*` and `Down*` fields.

2. You have to choose the options for processing and plotting the Up and Down reg. genes you entered.

   Options are place right below the `Down*` field. First one is `Colorscheme*`, which is a drop down menu in which you can choose the color scheme for the result heat map. In the next three options fields, you are provided with three different ways to assign which parts of the compendium to be used in the analysis. For example if you want to compare your own sample to *Hormone* samples in compendium, in the `Categories` section you have to only tick *Hormone* and then hit **Submit**. There are different categorical options in the `Accessions` and `Descriptions` fields.

2.2 "Submit csv" section

Basically the same as "Submit List of Genes" with two main differences:

- More than one sample could be entered and analyzed.
- Gene expression values are entered as input
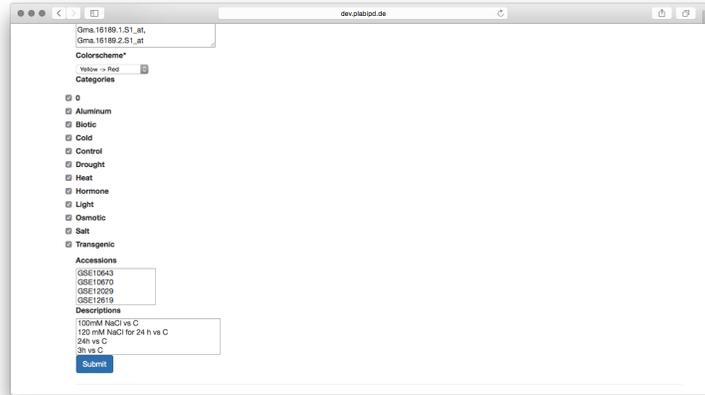
Figure 3: Gene list tab options

As the name suggests, this section gets a CSV file as the input. Opening this section, you encounter a page like figure 4, in which there are three input/option
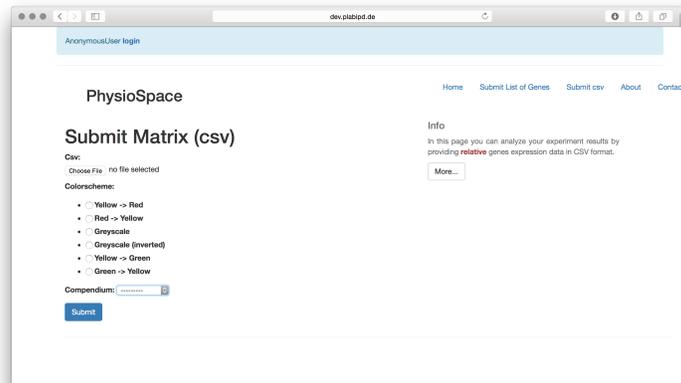
Figure 4: CSV section

fields.

The first one is **CSV** in which you `choose` and upload a CSV file of your experiment gene expressions. When making a CSV file to use as an input, remember that this tool expects the CSV to have **row** and **column names**.

- Column names are used as labels in the generated heat map of results.

- Correct Row names are very crucial because they are used to match genes between input and compendium so they must be the same type. For

example, ATH1 compendium is a compendium made of GEO data of ATH1 Affymetrix arabidopsis DNA array, so its row names are Affymetrix probe IDs, so when using this compendium the row names of your input CSV file should also be Affymetrix probe IDs.

After the **CSV** there are two options fields: in the first one **colorscheme** you can set the color scheme to use for the heat map showing the results.

And the last option **Compendium** gives you the option of choosing *in comparison to which reference you want to calculate you similarity scores for your experiment sample.* To put it simple: PhysioSpace compares your new experiment data to the existing, already available and reliable data set, **Compendium** is that data set.

After setting these three fields and hitting the **Submit** button, the heat map of results will be shown. If you're already a logged-in user, these results will be saved and can be accessed later from the "home" tab.

3 Sample

Here we provide a complete sample for using Plant PhysioSpace. For preparing the data (download, normalize, Differentially expressed gene calculation, ...), we used R (https://www.r-project.org) while any other software can be used as long as at the end the proper CSV file is generated. We decided to analyze a data set from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). We searched for two key words: "arabidopsis" and "stress" and chose the first returned result to analyze which was GSE22671 (http://www.ncbi.nlm.nih.gov /geo/query/acc.cgi?acc=GSE22671).

For starters, we're gonna download the data:

```
1  library(GEOquery) #Using GEOquery package to work with GEO
2  GSE22671 <- getGEO(GEO = "GSE22671",getGPL = F) #Downloading
3  GSE22671 <- GSE22671[[1]] #there is only one platform so it's
       easier to work with ExpressionSet data type rather than the
       List
```

GSE22671 is a light stress experiment with 9 samples in 3 conditions: `Control`, `Dark`, and `Light`. PhysioSpace only accepts **relative** a.k.a **Fold Change** values so we have to subtract each stress replicate from the corresponding Control:

```
1  relativGeneExpress <- matrix(NA,nrow = nrow(GSE22671),ncol = 6) #
       initializing
2  geneExpressionMatrix <- exprs(GSE22671) #just for easier use in the
       following loop
3  for(k in 4:9){
4      relativGeneExpress[,k-3] <- geneExpressionMatrix[,k]-
       geneExpressionMatrix[,ifelse(k<7,k-3,k-6)] #Relative gene
       expression calculation
5  }
6  rownames(relativGeneExpress) <- rownames(geneExpressionMatrix) #
       Setting rownames
7  colnames(relativGeneExpress) <- GSE22671@phenoData@data$
       characteristics_ch1[4:9]# Setting column names
```

Two important points to consider here:

- For calculating the Fold changes we subtracted the stressed sample gene expressions from Controls because downloaded data from GEO is already normalized and converted to log-scale. If your data is not in log-scale, you have to convert it yourself before calculating relative values.

- We just copied the row names from GSE22671 to our relativGeneExpress matrix because GSE22671 is done using ATH1 Affymetrix DNA array so the row names are already Affymetrix probe IDs. If in your experiment data, another type of gene IDs or names are used and you want to compare your data to a Affymetrix compendium in Plant PhysioSpace (as we want to do here), you have to convert your Gene names yourself, using biomaRt for example.

Now the relative values are ready to be processed with PhysioSpace, we just have to export our matrix into a CSV file:

```
write.csv(x = relativGeneExpress, file = "~/Desktop/GSE22671.csv",
    row.names = TRUE) #Don't forget the rownames!
```

An screen shot of the first few rows of the CSV file are provided in figure 5.

	stress: Dark	stress: Dark	stress: Dark	stress: Light	stress: Light	stress: Light
244901_at	0.028702708	0.25912843	-0.495100804	0.934782626	0.28846788	-0.127805575
244902_at	-0.252047197	0.281592508	-0.0791187999999998	0.810109138000001	0.0599351160000001	0.0087313719999974
244903_at	-0.357756299	0.521758218	-0.407802495	0.866282826	0.434182993999999	-0.236620749
244904_at	-0.281199244	-0.162428118	-0.0264200849999998	0.120658743	0.0563280769999999	-0.141074723
244905_at	0.0827298289999998	0.210606745	0.163845233	-0.0296681799999998	0.023972884	-0.135031585
244906_at	-0.149147911999999	-0.039973152	0.134787422	0	0.0679210220000002	0.1284964519999999
244907_at	-0.069650658	-0.160992068	0.0907055539999999	0.03420508	0.038630924	0.0881865549999996
244908_at	0	-0.134654654	0.442131693	-0.0860403830000003	-0.0706204410000004	0.0976020379999998
244909_at	-0.106208389	-0.0741640860000001	0	-0.0943962220000003	-0.0064042120000008	0
244910_s_at	0.230391243	0.124161501	0.062050224	-0.0721604930000002	0.0369849060000003	0.147433478
244911_at	-0.051782468999999	0	0.0067340650000029	0.243050988	0.042436565	0.058804776000001
244912_at	0.083956160999997	0.0072452539999898	-0.10307309500001	0.485245488	0.295708706999999	0.15771947800001
244913_at	0.370078637	-0.0415433969999998	0.077131413	-0.0395142079999999	0.484025247	0.157110898
244914_at	0.064080277000004	0.0916430740000003	0.18235481	0.56484969000001	0.332730349	-0.055220201
244915_s_at	-0.123494946	0.016119411	-0.0426329560000003	-0.315793106	0.119407074	-0.0125565519999995
244916_at	0.151096487999999	-0.107165304	-0.145598534	-0.0506284530000003	0.0777340209999999	-0.0271970130000003
244917_at	-0.00464811100000073	0.12387559	-0.282382818	0.035537418000001	0.303361294	0.0470811199999996
244918_at	0.11403475	0	-0.25819165	0.293883383	-0.121355324	-0.220186641
244919_at	-0.243378401999999	0.176396617	-0.681157252	-0.019573147	-0.657959527	-0.62251858200001
244920_s_at	-0.271815885999999	0.198228139000001	-0.0086873809999966	0.737976878000001	0.297580193	0.153376838
244921_s_at	0	0	-0.425177445	1.059236711	0.273253973999999	0.0519115250000004
244922_s_at	0.137256026	-0.0980561940000007	-0.498921273000001	0.441624083	0.235138184	0.13088122
244923_x_at	-0.326403474999999	0	-0.751546155	0.54165359800001	-0.195595066	-0.0678323049999996
244924_at	-0.156312262	0.30664933	-0.603527533999999	0.133497161	-0.304024938	-0.519761794
244925_at	-0.0298264039999996	-0.30327184	-0.00308392700000049	-0.217654576999999	-0.0284452929999999	0.225471341

Figure 5: Generated CSV file Screen shot

At the end we can use this CSV file we made in the "Submit csv" tab of PhysioSpace and calculate the similarity results.

References

[1] Lenz M, Schuldt BM, Müller F-J, Schuppert A. Physiospace: Relating gene expression experiments from heterogeneous sources using shared physiological processes. *PLoS ONE*, (8(10)), 2013. doi: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077627.

[2] Joint research center for computational biomedicine. URL `http://www.combine.rwth-aachen.de`.

[3] Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K. The atgenexpress global stress expression data set: protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses. *The plant journal: for cell and molecular biology*, (50):347–363, 2007. doi: http://www.ncbi.nlm.nih.gov/pubmed/17376166.