flowCHIC - Analyze flow cytometric data of complex microbial communities based on histogram images

Schumann, J., Koch, C., Fetzer, I., Müller, S.
Helmholtz Centre for Environmental Research - UFZ
Department of Environmental Microbiology

March 4, 2014

Abstract

CHIC means cytometric histogram image comparison and is a tool for evaluation of cytometric datasets nearly automatically and very fast. On the first hand it was developed for evaluation of microbial cytometric data but it can also be used for any other cytometric dataset. It is based on the main idea that histograms are converted into images thus lowering information but enabling diagnosing systems in an instant. Although cell specific information is given up in this way it allows, on the other hand, huge datasets to be evaluated within very short time frames. It is, therefore, a tool to visualize cell population or community dynamics. Right now, the main application is monitoring microbial dynamics in managed natural environments such as wastewater, biogas, or groundwater.

1 Introduction

The background of the CHIC tool is published in [Koch et al., 2013a]. The CHIC tool contains three parts. The first part provides information on cell distributions and their respective abundances in histograms, the second part transforms vector based histograms into bitmap images while the third part evaluates and compares these images. Only the third part of the CHIC tool uses R and is provided in this package. The first part of the CHIC tool directly uses cytometric raw data (usually .fcs files) that can be obtained from any cytometric analysis. As the later evaluation occurs on grey value comparison in a starting step arbitrary color keys of cytometric histograms are converted into equally spaced grey graduations. Here, [Koch et al., 2013a] provided a script for easy application in Summit 3.1 but other conventional vector based image software can also be used such as Power Point. Make sure that the same grey graduation is used for all samples in one application. The second part creates images with instrumental noise or bead signals eliminated by gate setting using Image J. This program is also used to compare all images by assessing presence/absence of pixel values of two images at a time [Koch et al., 2013a].
Now the third part starts using the script provided within this package. As an outcome nonmetric multidimensional scaling (NMDS) is performed giving information on how similar/dissimilar certain samples are or how the structure of populations or communities is changing over time. Examples are discussed in [Koch et al., 2014] and also given below. To find out the driving forces of population variation the rationale of abiotic parameters such as pH values, temperature, or concentrations of chemicals can be evaluated. By using gate information obtained from the CyBar tool [Schumann et al., 2014] influences coming from certain subpopulations or subcommunities can also be detected. It needs to be stated that CHIC requires datasets that are highly calibrated by using e.g. fuorescent beads. Every shift in the adjustment of a cytometer will be discovered using CHIC and will have an influence on the outcome. CHIC can be used for any cytometric dataset. The strength of the tool lies in its easy application, especially if big datasets needs to be evaluated and its nearly automatic, person-independent application. The tool provides trends in community or population dynamics or informs on stability of cell systems.

2 Overview

Creation of images from flow cytometric raw data The software Summit 3.1 is used for creating images of flow cytometric raw data.

Image analysis with ImageJ software The software ImageJ is used for analysing the images created by Summit 3.1.

Similarity calculation in R The similarities of the images are calculated using nonmetric multidimensional scaling (NMDS).

3 Details

3.1 Creation of images from flow cytometric raw data

Summit 3.1 is used for creating images of flow cytometric raw data. After the .fcs files of interest are opened the histogram properties are changed. The resolution is set to 128 and the color table is set for a grey value scale. The resulting images are saved as .bmp files in one folder. This procedure is described more detailed in [Koch et al., 2013a] and not a part of this package, by now.

3.2 Image analysis with ImageJ software

The images created by Summit 3.1 are processed using ImageJ. Using various macros the area of interest is cut from each picture manually, the overlap and XOR images are created and the areas and intensities are calculated. This procedure is described more detailed in [Koch et al., 2013a] and not a part of this package, by now.

3.3 Similarity calculation in R

The similarities found in the processed images are calculated using R. Therefore, a dissimilarity matrix is calculated based on the intensities of the XOR images. The results are shown as an NMDS plot. In addition, a cluster analysis is performed to reveal samples with high similarities. For additional information see [Koch et al., 2013a].

4 Reading the data

The data always have to be read in R. In R change the working directory to the path where the Results_overlaps and Results_XOR tables created by ImageJ are located and read the data (see example below). All backslashes "\" may be changed to slashes "/" when using Windows.

```
> # Type the path of the data into the quotes
> setwd("")
> # or in Windows use
> setwd(choose.dir(getwd(), "Choose the folder containing the data"))
> # Type the filename of the overlap data (including extension) into the quotes
> Results_overlaps<-read.table("Results_overlaps.txt",header=TRUE,sep="\t")
> # Type the filename of the XOR data (including extension) into the quotes
> Results_xor<-read.table("Results_xor.txt",header=TRUE,sep="\t")
```

5 CHIC

This package provides the third part of the CHIC analyzing procedure [Koch et al., 2013a] and is used for calculating the similarities found in the histogram images of cytometric data. This sections gives an overview of the function that is used in this package. The datasets that are attached to the package serve as examples for demonstration and the influence of different parameter settings on the results. The package has to be loaded in R.

```
> # Load package
> library(flowCHIC)
```

In the following the attached datasets `Results_overlaps` and `Results_xor` are used for demonstration.

The contents of these datasets looks as follows (excerpts):

```
> # Load dataset
> data(Results_overlaps)
> # Show data (first ten lines)
> Results_overlaps[1:10,]
```

```
                              Label   Area
1   overlap_Sample_1_&_Sample_2.bmp 116714
2   overlap_Sample_1_&_Sample_3.bmp 117420
3   overlap_Sample_1_&_Sample_4.bmp 118123
4   overlap_Sample_1_&_Sample_5.bmp 117999
5   overlap_Sample_1_&_Sample_6.bmp 116898
6   overlap_Sample_1_&_Sample_7.bmp 117459
7   overlap_Sample_1_&_Sample_8.bmp 116771
8   overlap_Sample_1_&_Sample_9.bmp 117150
9  overlap_Sample_1_&_Sample_10.bmp 116647
10 overlap_Sample_1_&_Sample_11.bmp 116895
```

```
> # Load dataset
> data(Results_xor)
> # Show data (first ten lines)
> Results_xor[1:10,]
```

```
                         Label  IntDen RawIntDen
1   xor_Sample_1_&_Sample_2.bmp 2343652   2343652
2   xor_Sample_1_&_Sample_3.bmp 2378996   2378996
3   xor_Sample_1_&_Sample_4.bmp 2468411   2468411
4   xor_Sample_1_&_Sample_5.bmp 2403830   2403830
5   xor_Sample_1_&_Sample_6.bmp 2415531   2415531
6   xor_Sample_1_&_Sample_7.bmp 2513456   2513456
7   xor_Sample_1_&_Sample_8.bmp 2806569   2806569
8   xor_Sample_1_&_Sample_9.bmp 2628781   2628781
9  xor_Sample_1_&_Sample_10.bmp 2550909   2550909
10 xor_Sample_1_&_Sample_11.bmp 2549358   2549358
```

A NMDS plot is displayed using the names of the samples as data point texts. In addition, a cluster dendrogram is calculated to reveal samples with high similarities. The order and the number of the samples is printed to the R console.

```
> chic(Results_overlaps,Results_xor)
```

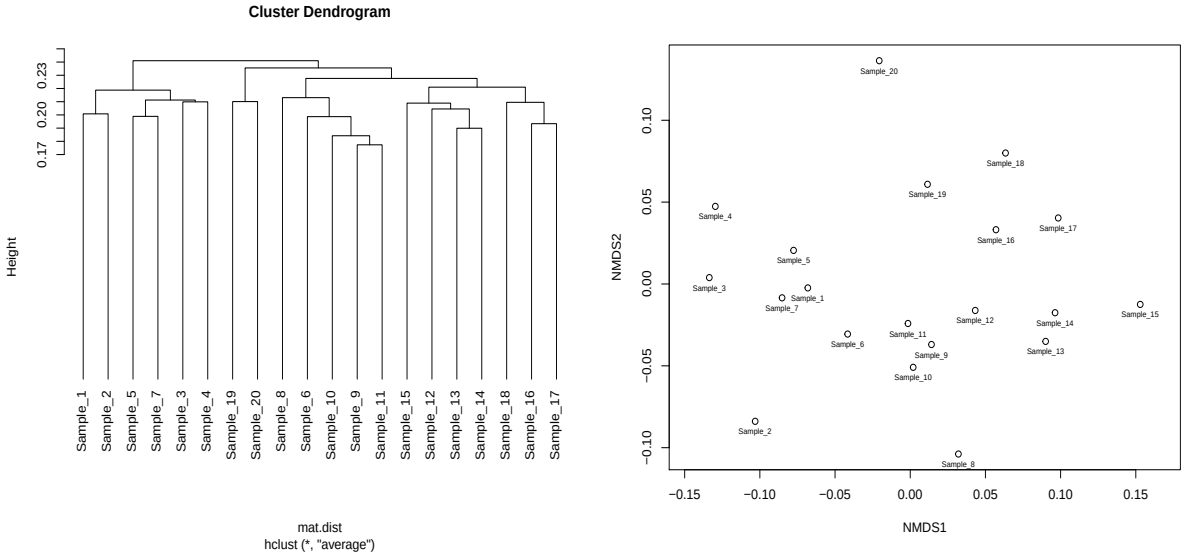

Figure 1: Cluster dendrogram and NMDS plot using default parameters

In addition, abiotic parameters can be added to the NMDS plot correlating with the differences found within the plot. It is important that the order and the number of the lines are identical to the order and the number of the samples printed in R. Otherwise, the output will be wrong since association of abiotic data with their samples will fail. The significance level p is set to 0.05 and the color used for plotting is set to `"magenta"` by default. If p is changed to 1 all abiotic parameters will be shown. In this example the attached dataset `Abiotic_data` is used for demonstration. The content looks as follows (excerpt):

```
> # Load dataset
> data(Abiotic_data)
> # Show data (first ten lines)
> Abiotic_data[1:10,]
```

```
   acetic propionic ibutter nbutter ivaleric
1    89.8       0.0     0.0     6.6      0.0
2    89.8       0.0     0.0     6.6      0.0
3    89.8       0.0     0.0     6.6      0.0
4    89.8       0.0     0.0     6.6      0.0
5    99.1       6.2     1.3     0.7      4.6
6    99.1       6.2     1.3     0.7      4.6
7    99.1       6.2     1.3     0.7      4.6
8    99.1       6.2     1.3     0.7      4.6
9    99.1       6.2     1.3     0.7      4.6
10   99.1       6.2     1.3     0.7      4.6
```

```
> chic(Results_overlaps,Results_xor,abiotic=Abiotic_data,show_cluster=FALSE)
```

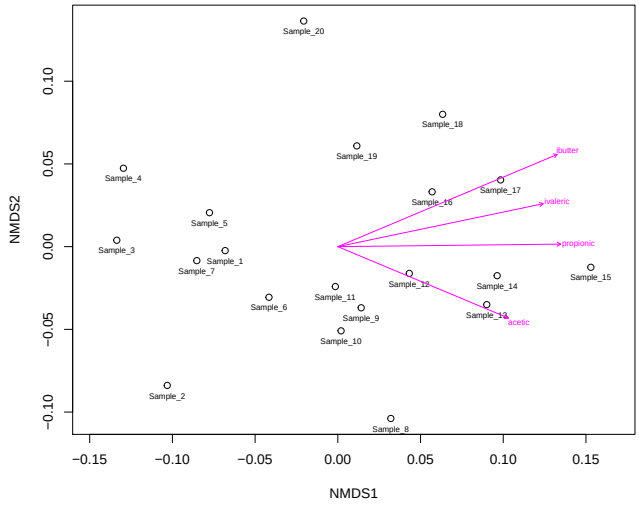

Figure 2: NMDS plot with abiotic parameters

The samples can be classified into groups manually. Therefore, every sample has to be assigned to one group. Again, it is important that the order and the number of the lines are identical to the order and the number of the samples printed in R. The file containing the group assignments can be read like described above. In this example 20 samples are assigned to three groups using the attached dataset `Groups`.

The content looks as follows (excerpt):

```
> # Load dataset
> data(Groups)
> Groups
```

	group
1	1
2	1
3	1
4	1
5	2
⋮	⋮
20	3

Table 1: Group assignments

This table describes the assignment of every sample to one group (three groups overall in this example). Every line (sample) is assigned to one value (1-25) so the same value means the same group. Up to 25 groups are possible (as only 25 plotting symbols are available). The samples can be grouped by e.g. treatment, day of sampling or personal preferences.

```
> chic(Results_overlaps,Results_xor,group=Groups,show_cluster=FALSE)
```

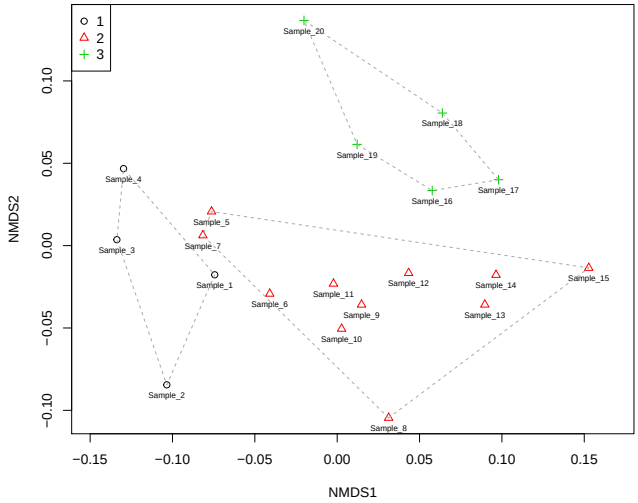

Figure 3: NMDS plot with groups

Plotting of the abiotic parameters and the groups can be combined in one plot.

```
> chic(Results_overlaps,Results_xor,group=Groups,abiotic=Abiotic_data,
+ show_cluster=FALSE,verbose=TRUE)
```

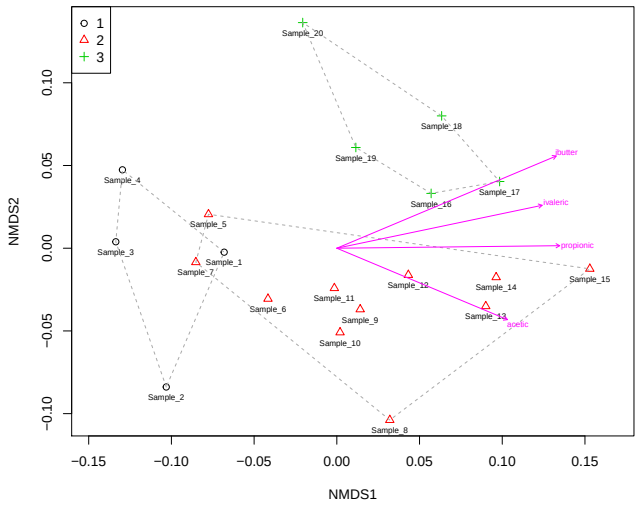

Figure 4: NMDS plot with abiotic parameters and groups

The p-value can be used to limit the display of abiotic parameters to the most significant ones. It is set to 0.05 per default. The color used for plotting these parameters is set to `"magenta"`. In the next example the p-value is changed to 0.01 and the color is changed to "darkred". Other possible colors are shown by typing colors() into R command line. In addition the legend is plotted to the top right corner of the plot.

```
> chic(Results_overlaps,Results_xor,group=Groups,legend_pos="topright",
+ abiotic=Abiotic_data,p.max=0.01,col_abiotic="darkred",show_cluster=FALSE)
```

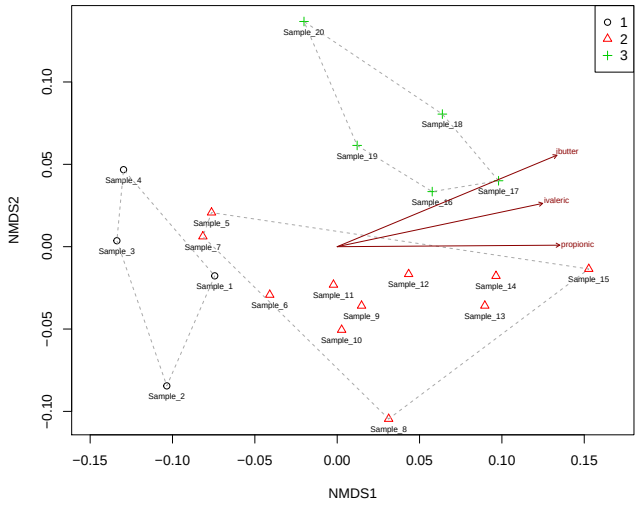

Figure 5: NMDS plot with less abiotic parameters and different color

Only three abiotic parameters (former: four, see Figure 4) are plotted due to the change of the maximal p-value.

Even though gating is not necessary for this analysis additional gate information can be added to the plot if they are available e.g. based on CyBar analysis [Koch et al., 2013b]. The attached dataset `Abiotic_gates` contains abiotic parameters as well as gate information (percental cell numbers of every gate for each sample).

The content looks as follows (excerpt):

```
> # Load dataset
> data(Abiotic_gates)
> # Show data (first ten lines)
> Abiotic_gates[1:10,]
```

	acetic	propionic	ibutter	nbutter	ivaleric	G1	G2	G3	G4	G5
1	89.8	0.0	0.0	6.6	0.0	14.81	5.53	2.93	2.73	19.68
2	89.8	0.0	0.0	6.6	0.0	14.40	5.43	3.21	2.95	18.27
3	89.8	0.0	0.0	6.6	0.0	15.83	5.10	2.75	2.73	20.41
4	89.8	0.0	0.0	6.6	0.0	16.60	5.34	2.61	2.76	18.39
5	99.1	6.2	1.3	0.7	4.6	15.21	5.25	2.67	2.60	19.19
6	99.1	6.2	1.3	0.7	4.6	14.52	5.58	2.88	2.89	20.18
7	99.1	6.2	1.3	0.7	4.6	15.51	5.21	2.69	3.08	21.18
8	99.1	6.2	1.3	0.7	4.6	12.53	6.38	3.20	2.62	19.89
9	99.1	6.2	1.3	0.7	4.6	13.63	6.85	3.51	2.86	18.51
10	99.1	6.2	1.3	0.7	4.6	13.61	6.77	3.40	2.91	18.48

```
> chic(Results_overlaps,Results_xor,group=Groups,abiotic=Abiotic_gates,
+ show_cluster=FALSE)
```

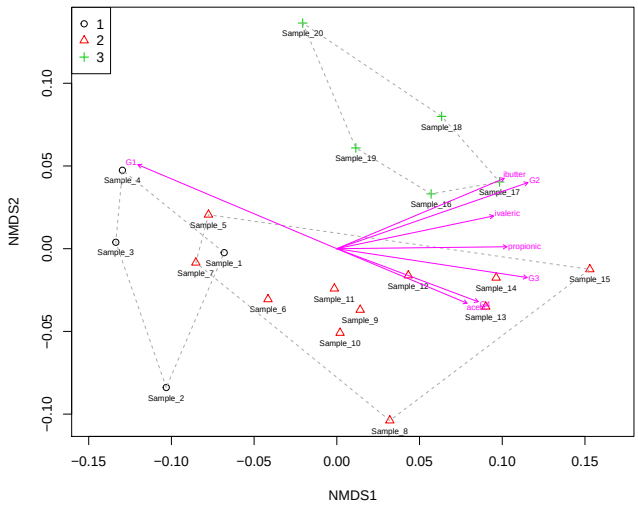

Figure 6: NMDS plot with groups, abiotic parameters and gate information

It is also possible to manually select the columns of the abiotic parameters that should be used for the analysis. This might be helpful if there is a high number of abiotic parameters but only the impact of a few of them is of interest. These columns can be selected using box brackets. Within the brackets the space in front of the comma defines the rows and the space behind the comma defines the columns to be selected. If one space is empty all rows/columns are selected. These spaces are filled using a concatenation of either numbers or strings. In this example only the columns 6-10 (G1-G5) are used for plotting.

```
> chic(Results_overlaps,Results_xor,group=Groups,
+ abiotic=Abiotic_gates[,6:ncol(Abiotic_gates)],show_cluster=FALSE)
```

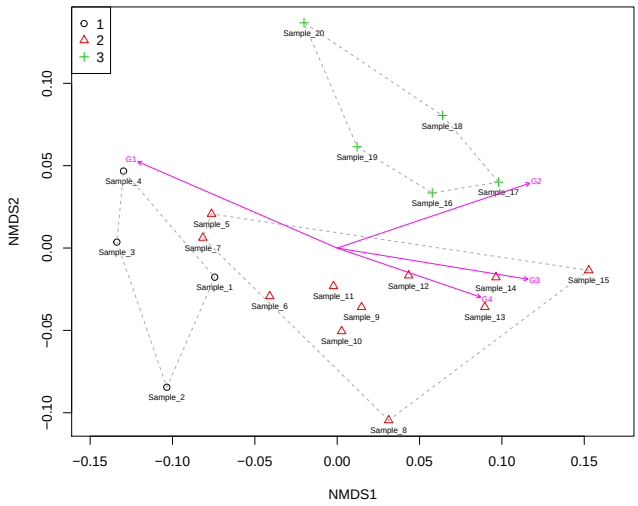

Figure 7: NMDS plot with selected parameters

Typing the names of the columns instead of their numbers will give the same result as shown in Figure 7.

```
> chic(Results_overlaps,Results_xor,group=Groups,
+ abiotic=Abiotic_gates[,c("G1","G2","G3","G4","G5")],show_cluster=FALSE)
```

6 Outlook

In the future this package will provide all three steps of the CHIC analyzing procedure. Users will no longer be forced to have Summit 3.1 installed on their computers due to the fact that Summit 3.1 is no freeware and outdated. They will also not need to execute ImageJ manually. Providing all three steps to this package will make this analysing procedure easier, faster and more powerful.

References

[Koch et al., 2013a] Koch, C., Fetzer, I., Harms, H., and Müller, S. (2013a). CHIC - an automated approach for the detection of dynamic variations in complex microbial communities. *Cytometry Part A*.

[Koch et al., 2013b] Koch, C., Fetzer, I., Schmidt, T., Harms, H., and Müller, S. (2013b). Monitoring functions in managed microbial systems by cytometric bar coding. *Environmental Science & Technology*, 47(3):1753–1760.

[Koch et al., 2014] Koch, C., Müller, S., Harms, H., and Harnisch, F. (2014). Microbiomes in bioenergy production: From analysis to management. *Current Opinion in Biotechnology*, 27:65–72.

[Schumann et al., 2014] Schumann, J., Koch, C., Günther, S., Fetzer, I., and Müller, S. (2014). *flowCyBar: Analyze flow cytometric data*. R package version 0.99.0.