

automatic_binner.r¹ manual

Alban Ramette
October 13, 2008

1. Preparing the input file

The GeneMapper output file containing the peak sizes, area and height can be copied to your favorite tabulation software.

	A	B	C	D
1	Sample	Size	Area	Height
2	A	5.05	385	4248
3	A	6.68	708	6109
4	A	504.57	6369	62531
5	A	518.68	86	2145
6	A	522.45	115	3180
7	A	525.54	120	2228
8	A	535.36	132	3215
9	A	535.97	135	2477
10	A	537.86	717	2485
11	A	599.38	66	1796
12	A	601.69	7346	107073
13	A	604.69	51	864
14	A	612.41	98	2168
15	A	654.33	5125	65429
16	A	681.66	3105	32818
17	A-1	4.74	817	9001
18	A-1	5.56	752	6618
19	A-1	7.27	285	2425
20	A-1	504.46	5482	58700
21	A-1	601.63	6948	100555
22	A-1	654.37	4223	57039
23	A-1	681.55	2689	25899
24	A-2	2.7	3907	31459
25	A-2	4.33	1001	12828
26	A-2	5.72	135	767

Copy the sample, size and area columns **only** to a text file (the height column is not needed). It is important to remove the lines that contain missing information. Column labels must be indicated. An example is given in [Data for binner.xls](#) in the “initial” sheet and in the corresponding [GeneMapperData1.txt](#).

2. Start R by clicking on the R icon

```
RGui - [R Console]
R version 2.7.0 (2008-04-22)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

3. Load the data into the R workspace

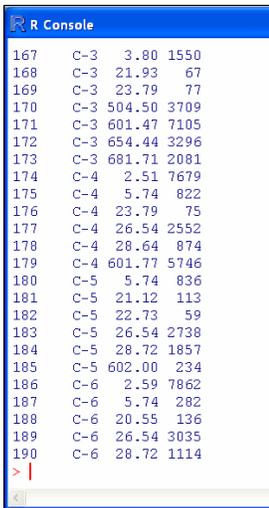
At the prompt (>), indicate in which directory you want to read and write the data (i.e. where you also put your .txt file. The directory should be created beforehand), and press enter. Note that quotes and \\ are used to indicate the path to the directory.

```
>setwd("c:\\R\\ARISA")
```

¹This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but **without any warranty**; without even the implied warranty of **merchantability** or **fitness for a particular purpose**. See the GNU General Public License for more details (Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307, USA)

Then, load the data into the object D by typing the following: (make sure to exactly type the dots, commas, and punctuation signs, as indicated and use " instead of ")

```
> D1=read.table("GeneMapperData1.txt",h=TRUE)
```



```

167 C-3 3.80 1550
168 C-3 21.93 67
169 C-3 23.79 77
170 C-3 504.50 3709
171 C-3 601.47 7105
172 C-3 654.44 3296
173 C-3 681.71 2081
174 C-4 2.51 7679
175 C-4 5.74 822
176 C-4 23.79 75
177 C-4 26.54 2552
178 C-4 28.64 874
179 C-4 601.77 5746
180 C-5 5.74 836
181 C-5 21.12 113
182 C-5 22.73 59
183 C-5 26.54 2738
184 C-5 28.72 1857
185 C-5 602.00 234
186 C-6 2.59 7862
187 C-6 5.74 282
188 C-6 20.55 136
189 C-6 26.54 3035
190 C-6 28.72 1114

```

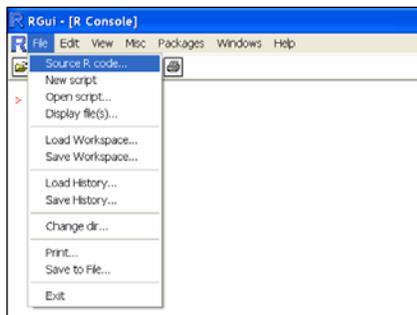
If you now type

```
>D1
```

You should see your data table appearing in the R console (if it is a big table, you will not see the first rows, but just the end of the table):

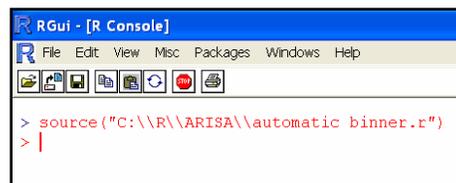
We are now ready to run the R script on the data stored in D.

4. Running the automatic binner script



In the menu bar, go to Source R code...

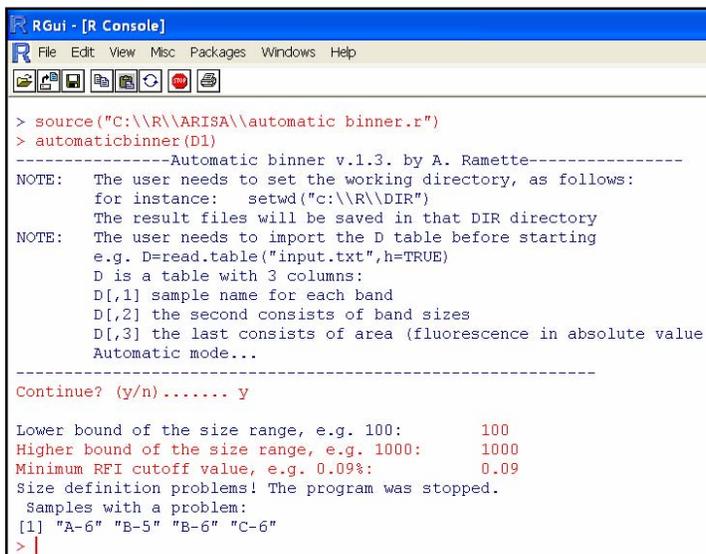
And indicate the location of your saved version of the [automatic binner.r](#) script.



```
> source("C:\\R\\ARISA\\automatic binner.r")
> |
```

```
>automaticbinner(D1)
```

The script starts by indicating some basic information about the version, expected data format and ask you if you want to proceed. Type "y" (without the quotes), if the table D corresponds to the description provided, otherwise type "n" and see the points above.



```

> source("C:\\R\\ARISA\\automatic binner.r")
> automaticbinner(D1)
-----Automatic binner v.1.3. by A. Ramette-----
NOTE: The user needs to set the working directory, as follows:
for instance: setwd("c:\\R\\DIR")
The result files will be saved in that DIR directory
NOTE: The user needs to import the D table before starting
e.g. D=read.table("input.txt",h=TRUE)
D is a table with 3 columns:
D[,1] sample name for each band
D[,2] the second consists of band sizes
D[,3] the last consists of area (fluorescence in absolute value
Automatic mode...
-----
Continue? (y/n)..... y

Lower bound of the size range, e.g. 100: 100
Higher bound of the size range, e.g. 1000: 1000
Minimum RFI cutoff value, e.g. 0.09%: 0.09
Size definition problems! The program was stopped.
Samples with a problem:
[1] "A-6" "B-5" "B-6" "C-6"
> |

```

For this example, we can use the following parameters:

- Smallest band size of the range 100
- Largest band size of the range 1000
- Minimum RFI cutoff of RFI 0.09

After few seconds, you should see the following message in the R console:

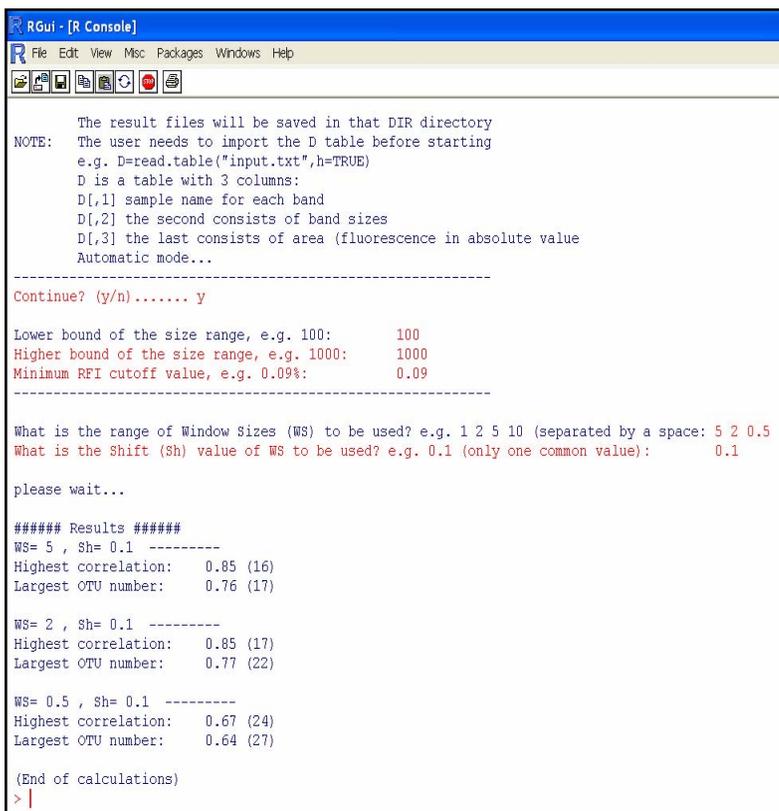
In this example, the script detected that for some samples, the highest peak size was not fitting in the predefined size range (e.g. for A-6 the largest peak was 28.8 bp while the selected range was 100-1000 bp). Those samples must then be manually removed from the data (i.e. you need to go back to point 1). In this tutorial, the corrected files are found in [Data for binner.xls](#) (“corrected” sheet) and the corresponding [GeneMapperData2.txt](#).

Reimport the data to the R workspace:

```
>D2=read.table("GeneMapperData2.txt",h=TRUE)
```

And run the script again:

```
>automaticbinner(D2)
```



```
RGui - [R Console]
File Edit View Misc Packages Windows Help

The result files will be saved in that DIR directory
NOTE: The user needs to import the D table before starting
e.g. D=read.table("input.txt",h=TRUE)
D is a table with 3 columns:
D[,1] sample name for each band
D[,2] the second consists of band sizes
D[,3] the last consists of area (fluorescence in absolute value)
Automatic mode...

-----
Continue? (y/n)..... y

Lower bound of the size range, e.g. 100:      100
Higher bound of the size range, e.g. 1000:    1000
Minimum RFI cutoff value, e.g. 0.09%:        0.09
-----

What is the range of Window Sizes (WS) to be used? e.g. 1 2 5 10 (separated by a space: 5 2 0.5
What is the Shift (Sh) value of WS to be used? e.g. 0.1 (only one common value): 0.1

please wait...

##### Results #####
WS= 5 , Sh= 0.1 -----
Highest correlation: 0.85 (16)
Largest OTU number: 0.76 (17)

WS= 2 , Sh= 0.1 -----
Highest correlation: 0.85 (17)
Largest OTU number: 0.77 (22)

WS= 0.5 , Sh= 0.1 -----
Highest correlation: 0.67 (24)
Largest OTU number: 0.64 (27)

(End of calculations)
>|
```

Again you may try the following parameters:

- Smallest band size of the range 100
- Largest band size of the range 1000
- Minimum RFI cutoff of RFI 0.09
- Window sizes to be screened: 5 2 0.5
(i.e. 5, 2, 0.5 bp)
- Shift value (Sh): 0.1
(only one value is allowed)

You should now obtain the following output:

This time, the script did not stop because no size problems were encountered and the calculations were done.

(Note that the script will also send an error message and stop if the RFI cutoff value is set too high. In the latter case, it would remove too many peaks for calculations to be correctly performed).

5. Analyzing the results

The console above indicates, for each window size, the highest correlation values among samples and the respective number of OTUs in parentheses. The first line corresponds to a focus on the highest correlation value among all calculated bin frames, whereas the second line reports the results with a focus on the largest OTU number. This way, each user can decide to use either highest correlation or highest OTU number to choose the corresponding WS value. Typically, a compromise between high correlation value and high number of OTUs has to be found. Note that the program does not return the calculations for the individual bin frames and the [interactive_binner.R](#) script should then be used with the chosen WS and Sh values. The console results can be saved by left-mouse selecting the text and copying by right-mouse clicking.

How to cite the script?

Ramette, A. (2008) Quantitative molecular community fingerprinting for estimating the abundance of operational taxonomic units in natural microbial communities. *submitted*