An Introduction to R

Antonio Mora

Institutt for molekylær biovitenskap --University of Oslo

and

The Biotechnology Centre of Oslo

MBV-INF 4410

Oslo, 8th September 2009

-- What is R ??

- A «language and environment for statistical computing and graphics» (previously known as «S»)
- Free, open-source, extended through packages.
- Well documented:

```
> help.start( )
> ?mean
> demo(graphics)
> vignette( )
```

- Last version: R 2.9.2 (August 24th, 2009)

-- R Syntax

- Syntax is similar to most scientific software.
- Arithmetic operators:

```
> 5 + 3
[1] 8
> 2 * 4 - 5
[1] 3
```

- Assigning values to **variables** (No declaration):

```
> a <- 5
> a <- 5 + 2
> a = 8 / 2
> a
[1] 4
```

- List all variables in workspace:

```
> ls( )
```

-- R Syntax (2)

- Vectors and lists:

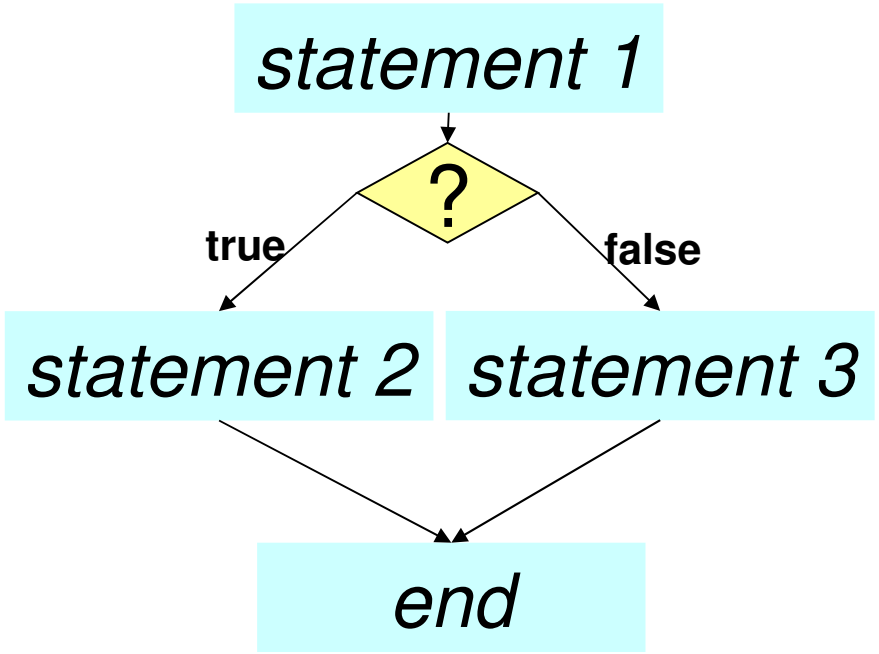
```
> a = c(2,3,4)
> a
[1] 2 3 4
> a[2]
[1] 3
> c = c(5,6)
> d = c(a,c)
> d
[1] 2 3 4 5 6
> d[1:4]
[1] 2 3 4 5
> l = list(1,2,«hi»)
> m = list(a=1,b=2,d=«hi»)
> m$b
[1] 2
```

- Printing out:

```
> print(m)
```

-- R Syntax (3)

- Conditions ('If') and Loops ('For') structures:




```
for (var=INITIAL ASSIGNMENT;
var=CONDITION; INCREMENT) {
    STATEMENT1;
    STATEMENT2;
    STATEMENT3;
    ...
}
```

```
> if (1>0) print(«hello»)
> for (i in 1:5) {print(i)}
> for (i in list(«a»,«b»,TRUE)) {print(i)}
```

-- R Syntax (4)

- Save and Load entire workspace or single variables:

```
> save(file=«variabs.Rdata»)
> load(file=«variabs.Rdata»)
> write(x,«scandata.txt»)
> p = scan(«scandata.txt»)
```

- Functions: Container of a group of statements. They are created in order to not to repeat the same group of statements every time.

```
> suma <- function(a,b) {a+b}
> suma(1,2)
[1] 3
```

- Libraries: Groups of functions with similar goals.

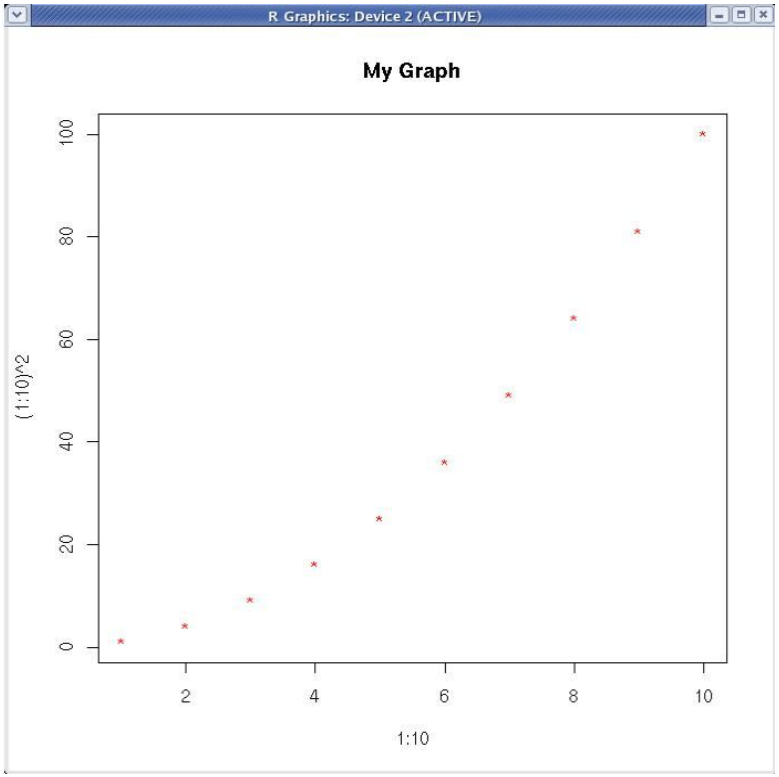
- Command Line execution and Scripts:

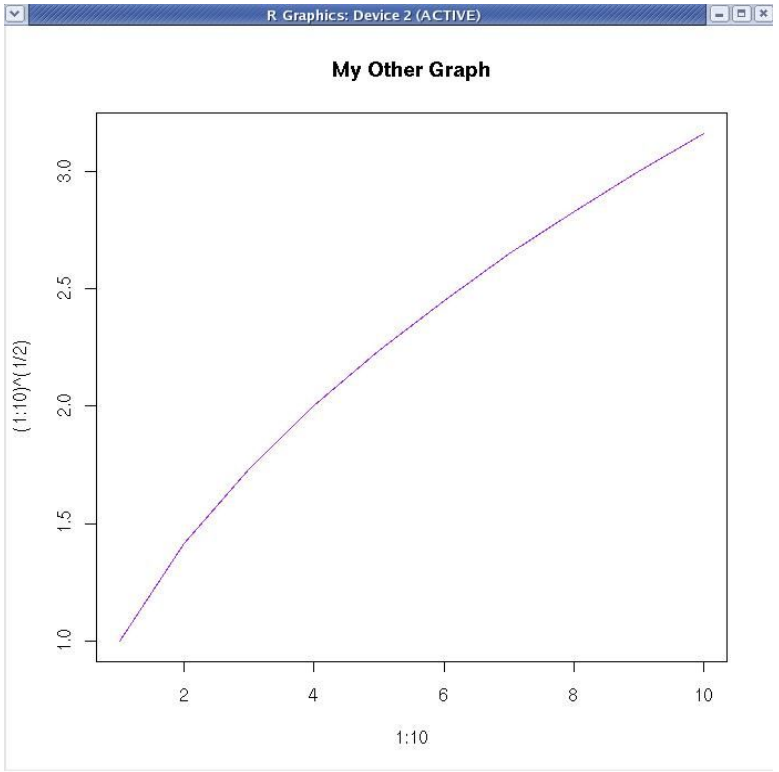
```
> source(«myProgram.R»)
> source(«http://www.bioconductor.org/getBioC.R»)
```

-- R Syntax (5)

- Plotting:

```
> plot(x=1:10)
> plot(x=1:10,y=(1:10)^2)
> plot(x=1:10, y=(1:10)^2, type=«p», col=«red»,
main=«My Graph», pch=«*»)
> for (i in 1:10) {x11( ); plot(i:10)}
```





-- R Syntax (6)

- Statistical functions:

```
> x = rpois(20,2)
> x = rnorm(5,1,2)
> mean(x)
> var(x)
> sum(x)
> cumsum(x)
```

- Reading text:

```
> write.table(mat,«mytable.txt»)
> c = read.table(«mytable.txt»)
```

-- R Syntax (7)

- Regular expressions --grep:

```
> grep("[a-z]", letters)
> txt <- c("arm","foot","lefroo", "bafoobar",
"gorilla","armadillo","far","tate")
> i <- grep("foo",txt)
> i=grep("^l.*",txt)
> i=grep("^f[oa]",txt)
> i=grep(".*o$",txt)
> i=grep("^...o{2}b..$",txt)
```

- Regular expressions --substitute:

```
> txt <- c("The", "licenses", "for", "most",
"software", "are", "designed", "to", "take",
"away", "your", "freedom", "to", "share", "and",
"change", "it.", "", "By", "contrast,", "the",
"GNU", "General", "Public", "License", "is",
"intended", "to", "guarantee", "your",
"freedom")
> ot <- sub("[b-e]",".", txt)
> ot2 <- sub("[b-e]{2}","..", txt)
```

-- What is BioConductor ??

- R package (group of packages/libraries) for Molecular Biology, Bioinformatics and Systems Biology
- Main categories:

  Annotation: GO, Pathways (KEGGSOAP), etc...

  AssayDomains: CellBasedAssays, ChIPchip, CopyNumberVariants, CpGIsland, DNACopyNumber, DNAMethylation, ExonArray, GeneExpression, GeneticVariability, SNP, Transcription...

  AssayTechnologies: Microarray, MassSpectrometry, SAGE, FlowCytometry, Sequencing, HighThroughputSequencing...

  BiologicalDomains: CellBiology, Genetics, Proteomics

  Infrastructure: DataImport, DataRepresentation, GraphsAndNetworks, GUI, Visualization

  Bioinformatics: Clustering, Classification, MultipleComparisons, QualityControl...
- Last version: Bioconductor 2.4 (21st April, 2009)

-- Installing BioConductor

- To get the standard distribution:

```
> source("http://www.bioconductor.org/getBioC.R")
> getBioC( )
```

- To get additional libraries (not included on standard distribution):

```
> source("http://www.bioconductor.org/getBioC.R")
> getBioC(c("graph", "RBGL", "GOstats", "ontoTools",
"ppiStats", "XML"))
```

-- Last thougths on R

- R is a scientific language, perhaps not as powerful as MatLab or Mathematica, but **continuously improved** and **extendible through packages**. And it is **free**.
- BioConductor contains **many Bioinformatics-related libraries**, covering a wide spectrum of applications, and a working community around them. The same is true for similar projects such as BioPerl or BioPython.
- Being a language thougth for statistics, it is recommended **when requiring statistical computation**.

-- References:

[1] BIOCONDUCTOR website, http://www.bioconductor.org/, 2007

[2] R website, http://www.r-project.org/, 2007

Now, let's start the lab !!