Visualizing statistics in R
(Part 1: Introduction)

Rita Giordano

PSI/SLS

Villigen, Switzerland

Quintiles

Strasbourg, France

OUTLINE

- R
- Packages
- How R works
- R programming
- Produce graphics for publications

Why R?

- R is FREE!!!!
- R is a language
- A flexible statistical toolkit
- R runs on a wide array of platform
- R gives you unlimited possibility to analyze your data.

How to run R

- Linux, Widows, Mac
- From command line
- R http://www.r-project.org
- Rstudio http://www.rstudio.com
- Rpy: http://rpy.sourceforge.net/rpy2.html

R packages

The packages for R can be downloaded from CRAN website:

http://cran.r-projects.org/web/packages

A packages is a collection of R function, data and compiled code. These are stored in the folder library

> library() gives you the path where are all your packages.

The standard packages included in R are:

base, datasets , graphics, stats, methods, utilis, grdevices

Package in crystallography

DISP: Diffraction Image Statistics Package
http://code.google.com/p/disp/

Waterman & Evans. Estimation of error in diffraction data measured by CCD area detectors. Journal of Applied Crystallography 43(6), 2010.

If you need a version for Mac, including function to read PILATUS detector data, please contact me.

Package for structural biological analysis

Bio3d http://thegrantlab.org/bio3d/

Contains utility for analysis of protein structure, sequence and trajectory data.

Bio3D: An R package for the comparative analysis of protein structures. Grant, Rodrigues, ElSawy, McCammon, Caves, (2006) *Bioinformatics* ***22, 2695-2696***

Graphic package: ggplot2

ggplot2 http://ggplot2.org

ggplot2 philosophy:

"Instead of spending time making your graph look pretty, you can focus on creating a graph that bests reveals the messages in your data."

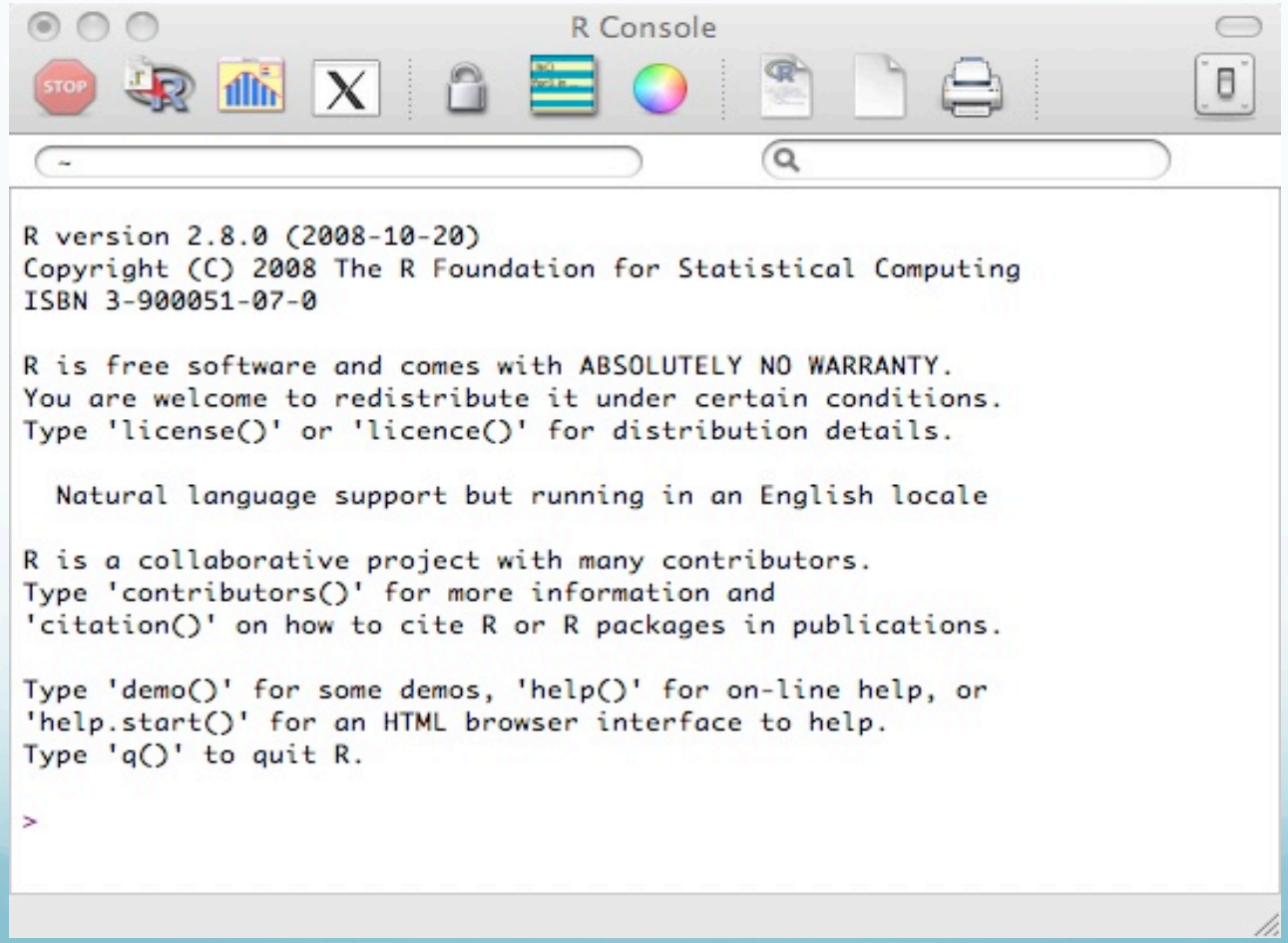
Interface between R and Python: RPy2

- http://rpy.sourceforge.net/rpy2.html

RPy2 is an interface between Python and R language. It can manage R objects and execute R function in a Python environment.

How work R?

R is an interpreted language accessible from command line.



Example of R session

Simple sum as a calculator

```
> 1+1
[1] 2
```

Getting help:

```
> help.start()                    # general help
> ?mean                           # help for mean function
```

R use the symbol <- for assignment =

```
>  X<-1+1                         # this will create an object named X
> X
[1] 2
```

Function to manage R workspace

R workspace is the current working R environments

> getwd()	list the directory
> setwd("path")	change the current directory
> ls()	list the objects
> rm(objectlist)	remove object
> history()	display history
> q()	quit R session

Data structures

R has a wide variety of object to hold data:

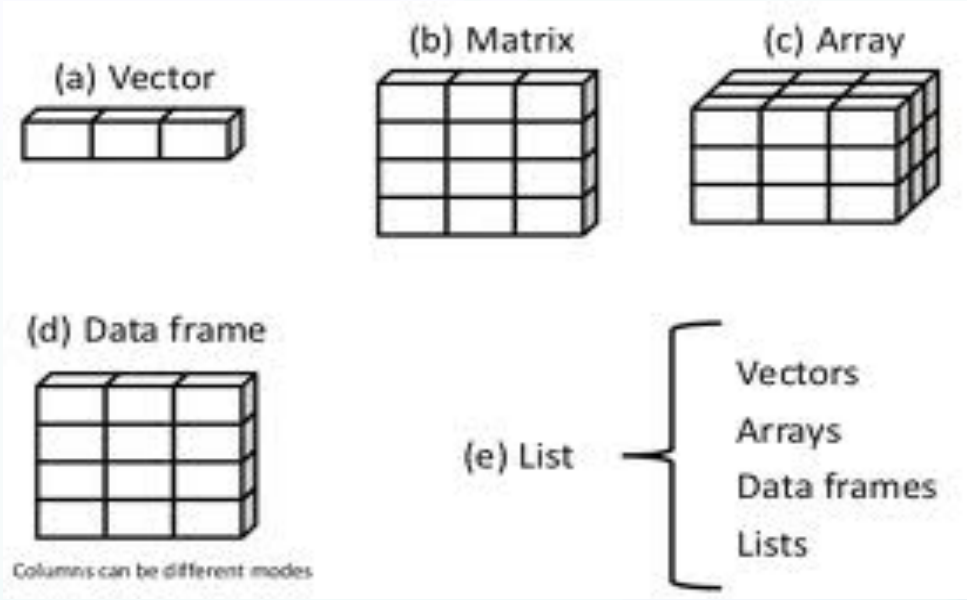
a) Vector 1-dimensional

b) Matrix 2-dimensional

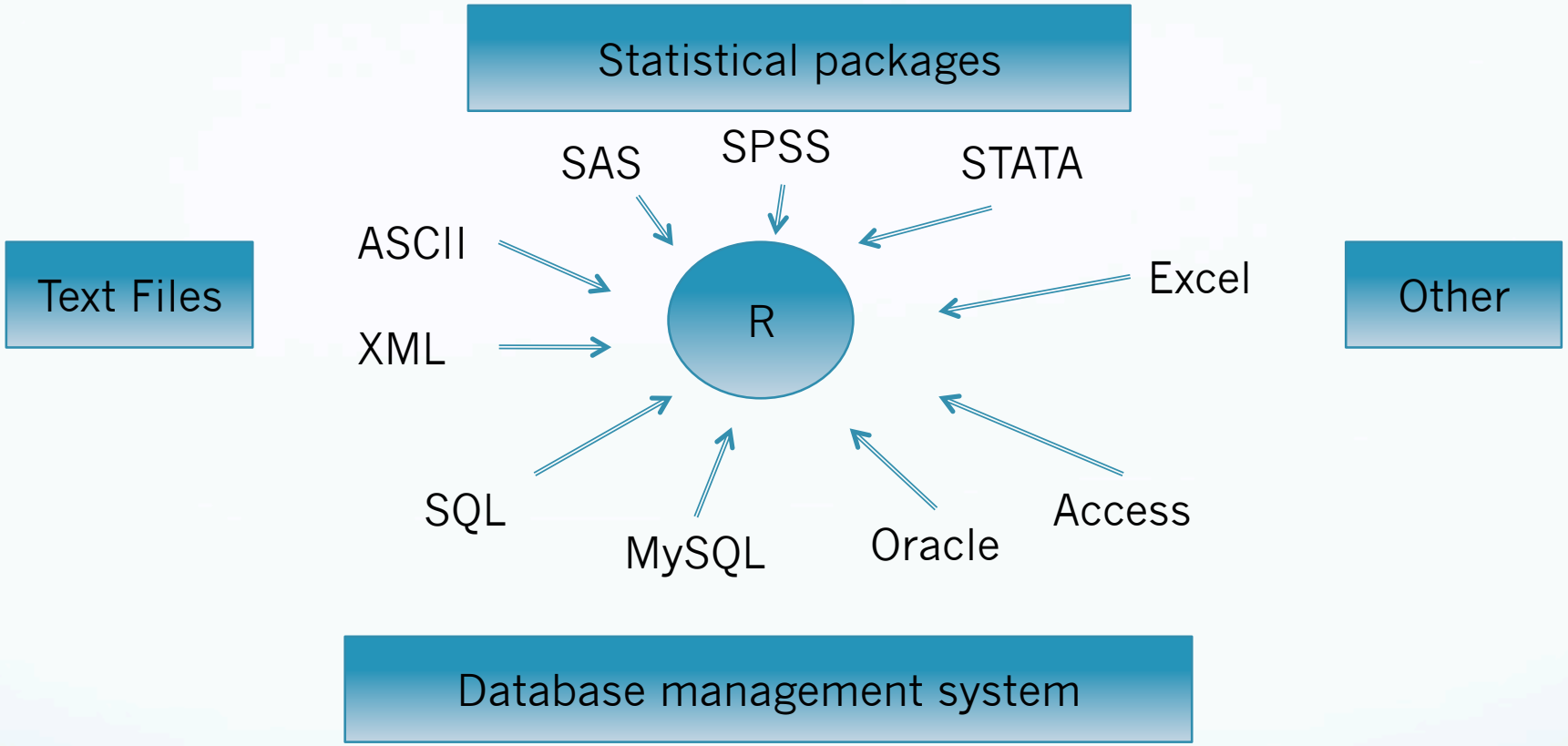
c) Array > 2-dimensional

d) Data Frame 2-dimensional (character and numeric)

e) List ordered collection of objects



Data input



Text file:

Data <- read.table('X-ray_parameters.dat')

Input file csv:

csv<- read.csv("x-ray.csv", header=True, sep=',')

Read data from X-Ray diffraction

- DISP package

  1) read data from CCD detector

     readImage('x-ray.img')

  2) Read data from PILATUS detector

     readCBF('x-ray.cbf')

Read data from X-Ray diffraction

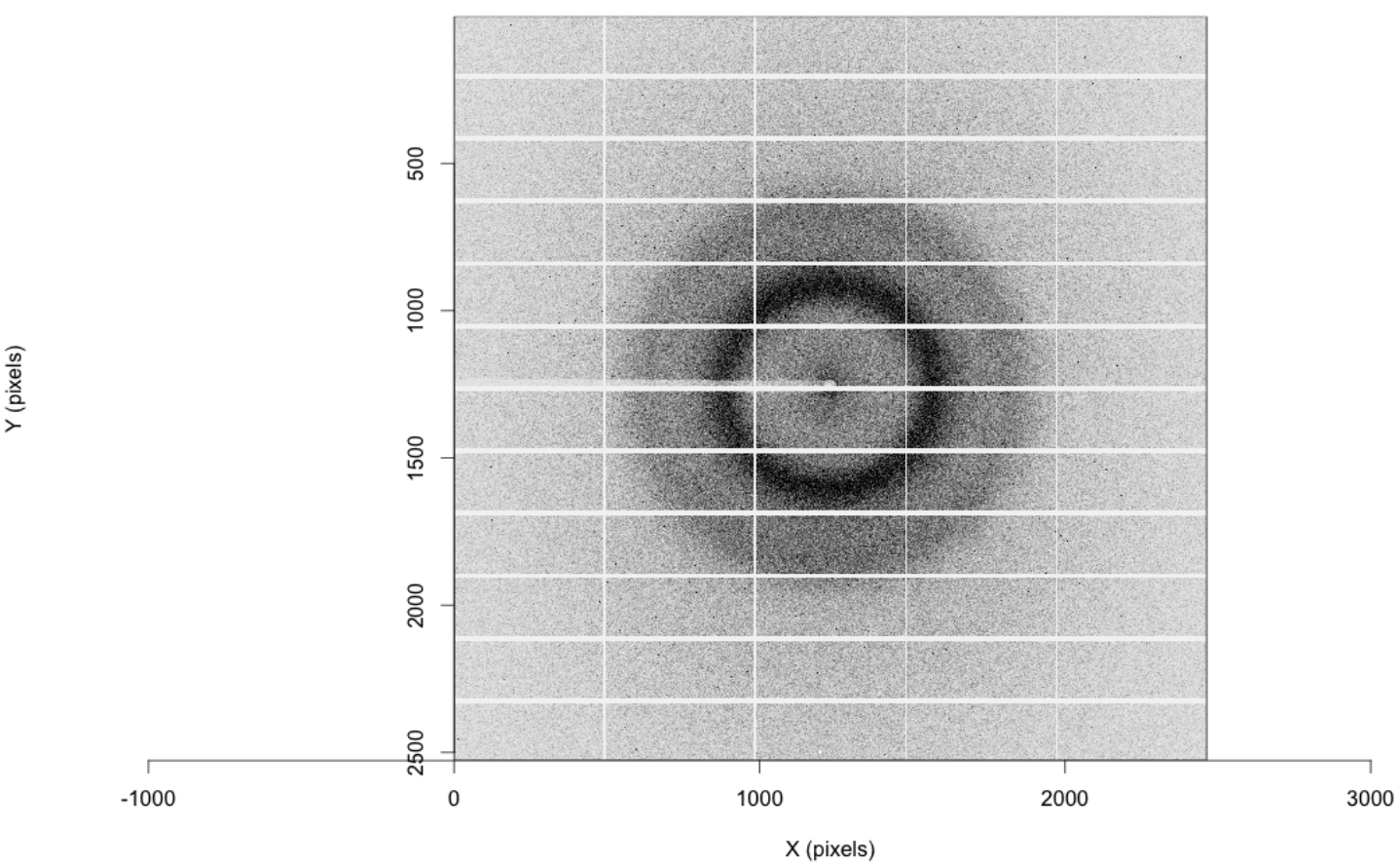
- DISP package

  1) rea

     re

  2) Re

     re



Read data from X-Ray diffraction data processing

- XDS files: XDS_ASCII.HKL, INTEGRATE.HKL, XSCALE.ahkl are already ASCII file, in the function read.table() we have only specify to not read the header.

  hkl<-read.table("XDS_ASCII.HKL",skip=31)

- MTZ files: NOT ASCII FILES!!!!

  before read with R we have to convert to ASCII using the CCP4 program mtz2various

  DISP also read mtzfile: readMTZ("data.mtz")

Read data from pdb coordinates files

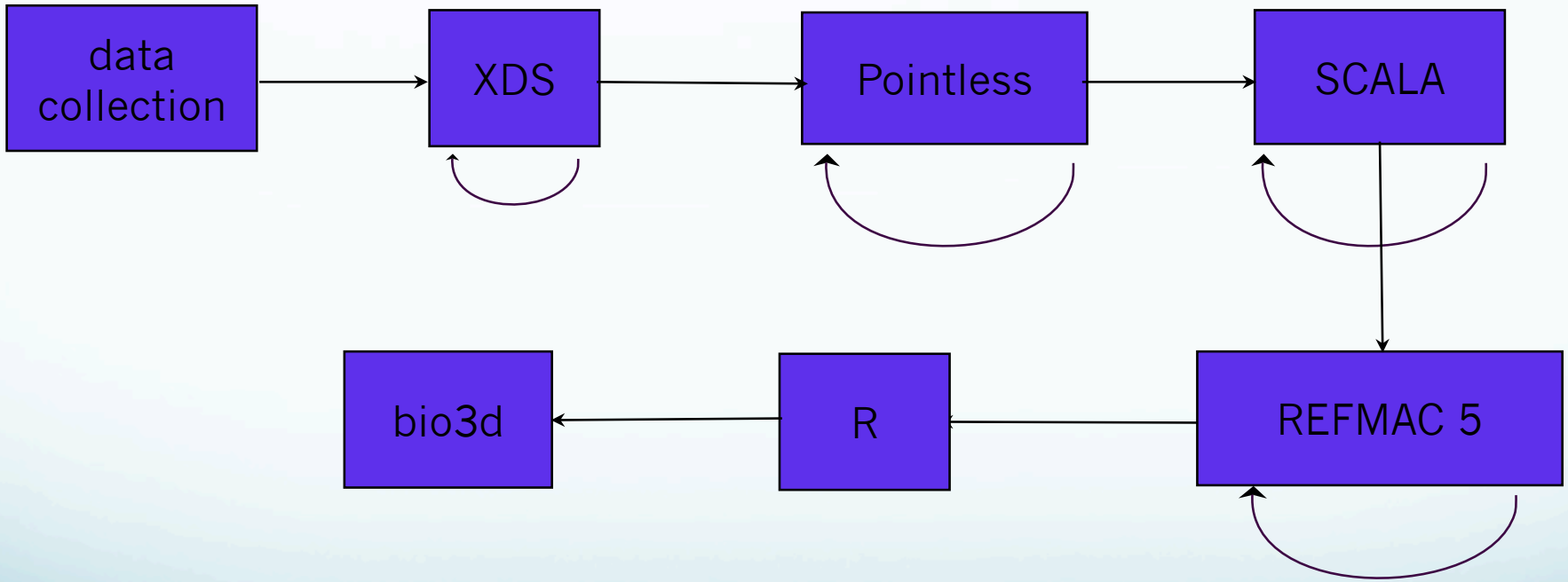
- Package BIO3D. After the refinement possible to read, modify and write pdb files.

  read.pdb('protein.pdb')

  this function return the following value:

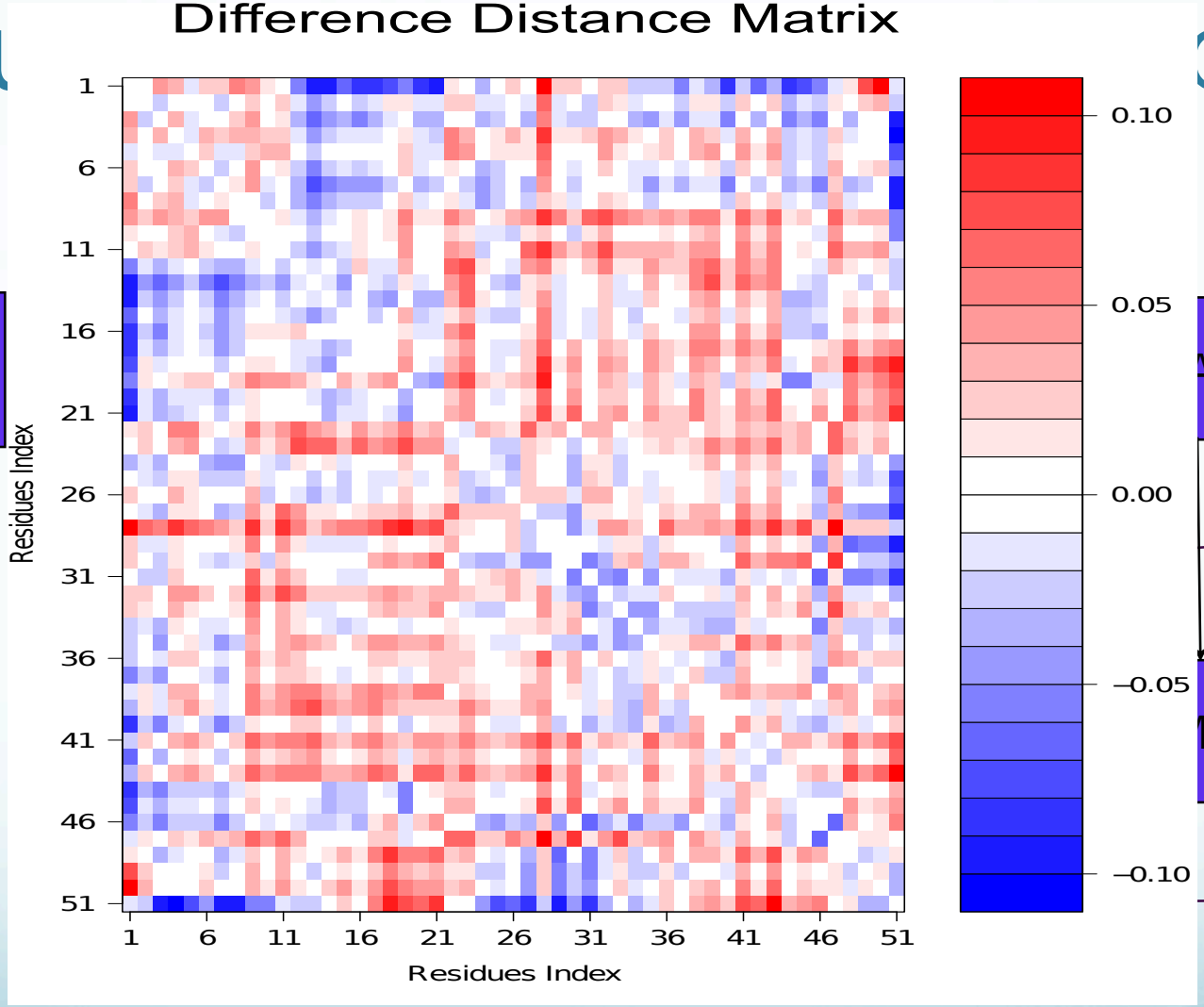
  Atom position, B-factor value, Occupancy, etc..

Structure analysis with bio3d



Stru o3d

data collection



ALA

MAC 5

Mathematical function

Function	Meaning
log(x)	log to base e of x
exp(x)	antilog of x (e^x)
log(x,n)	log to base n of x
log10(x)	log to base 10 of x
sqrt(x)	square root of x
factorial(x)	$x!$
choose(n,x)	binomial coefficients $n!/(x!\ (n-x)!)$
gamma(x)	$\Gamma(x)$, for real x $(x-1)!$, for integer x
lgamma(x)	natural log of $\Gamma(x)$
floor(x)	greatest integer $< x$
ceiling(x)	smallest integer $> x$
trunc(x)	closest integer to x between x and 0 trunc(1.5) = 1, trunc(-1.5) = −1 trunc is like floor for positive values and like ceiling for negative values
round(x, digits=0)	round the value of x to an integer
signif(x, digits=6)	give x to 6 digits in scientific notation
runif(n)	generates n random numbers between 0 and 1 from a uniform distribution
cos(x)	cosine of x in radians
sin(x)	sine of x in radians
tan(x)	tangent of x in radians
acos(x), asin(x), atan(x)	inverse trigonometric transformations of real or complex numbers
acosh(x), asinh(x), atanh(x)	inverse hyperbolic trigonometric transformations of real or complex numbers
abs(x)	the absolute value of x, ignoring the minus sign if there is one

Writing R function

function(argument list) expression

e.g Skewness:

$$m = \frac{\sum (y - \bar{y})^3}{n}$$

$$s = sd(y)^3$$

$$skew = \frac{m}{s}$$

```
skew<-function(x) {
        m3<-sum((x-mean(x))^3)/length(x)
        s3<-sqrt(var(x))^3
        m3/s3}
```

R programming

for(condition) {expression}

```
> for (i in 1:10) {x[i]=i}
> x
 [1]  1  2  3  4  5  6  7  8  9 10
```

If(condition) {expression}

```
> if (x[i]>1){y=x+1}
> y
 [1]  2  3  4  5  6  7  8  9 10 11
```

R Programming

Ifelse(test,yes,no)

```
> x<-c(6:-4)

> x

[1]  6  5  4  3  2  1  0 -1 -2 -3 -4

> sqrt(x)

 [1] 2.449490 2.236068 2.000000 1.732051 1.414214 1.000000
0.000000      NaN      NaN      NaN          NaN
Warning message: In sqrt(x) : NaNs produced

> sqrt(ifelse(x>=0,x,NA))

[1] 2.449490 2.236068 2.000000 1.732051 1.414214 1.000000
  0.000000       NA       NA       NA  NA
```

Statistical methods with R

(more details during the tutorial)

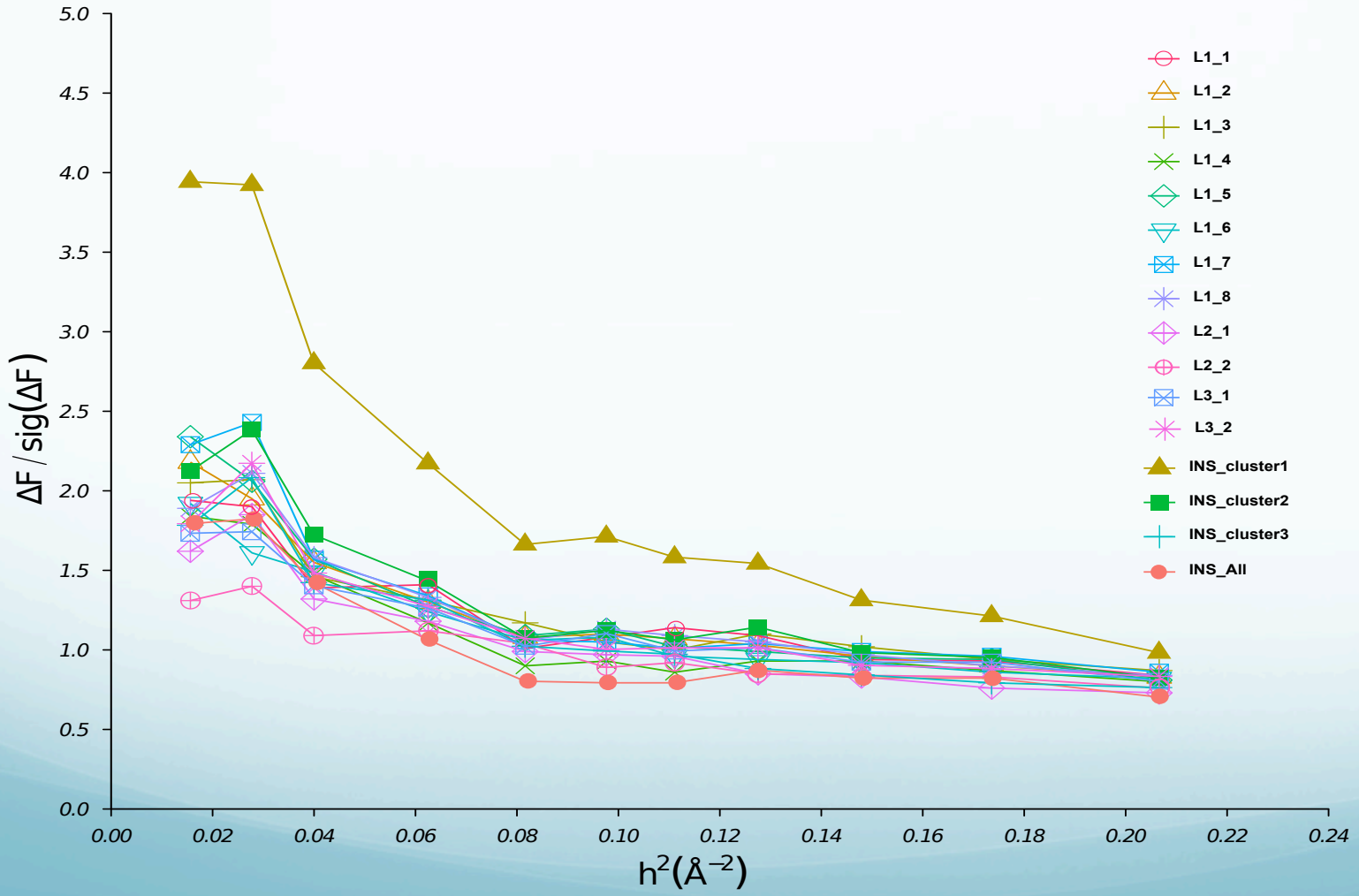
- ANOVA (Analysis if Variance)
- Multivariate statistical analysis (tutorial bio3d)
  - Cluster analysis
  - Principal components analysis
- Fit

Programming with RPy2 using python environment

```
import rpy2
import rpy2.robjects as robjects
import numpy as np
from numpy import *
import rpy2.robjects.numpy2ri as np2r
from rpy2.robjects.packages import importr

# Define robjects
r = robjects.r
stats = importr('stats')
```

How to prepare plot for publication



```
bdataframe<-data.frame(data)

b=qplot(B_fac, data=b_df, geom=point",xlab,ylab)
## Adding layer to the plot

b + geom_points( aes(x,y), colour = "black”) +

theme_bw() + scale_x_continuous(breaks=c(‘your_range’))+

opts(axis.text.x = theme_text(colour = 'black', angle = 0, size = 8, face = 'italic'))+
opts(axis.text.y = theme_text(colour = 'black', angle = 0, size = 8, face = 'italic'))+
opts(axis.title.x = theme_text(colour = 'black', angle = 0, size = 9.5, hjust = 0.5, vjust
= 0, face = 'italic'))+
opts(axis.title.y = theme_text(colour = 'black', angle = 90, size = 9.5,  face = 'italic'))
+
opts(plot.title=theme_text(size = 12))+
scale_y_continuous( expand=c(0,0),limits=c(0,50)) +
scale_x_continuous( expand=c(0,0), limits=c(17,22), breaks=seq(17,22,0.5))+
opts(panel.grid.minor = theme_blank(), panel.grid.major =
theme_blank(),legend.position = c(0.9,0.78),
            panel.border = theme_border(c("left","bottom")))+
opts(panel.grid.minor = theme_blank(), panel.grid.major = theme_blank())
```

Reference and useful website

- http://www.inside-r.org/
- http://www.r-bloggers.com
- http://www.statmethods.net
- http://www.ccp4.ac.uk/newsletters/newsletter49/articles/RforCCP4.pdf Article by James Foadi.
- Robert I. Kabacoff "R in Action". Manning
- Michael J Crawley "The R book". Wiley

Acknowledgements

MX group, SLS PSI
May Marsh, SLS Villig
Ezequiel Panepucci, SLS
Meitian Wang, SLS
Giovanna Miritello, Telefonica Madrid
Natalia Treissard, Quintiles Strasbourg
Sean McSweneey, ESRF Grenoble

Thank you for your attention

Advertisement time

- LOOKING FOR:

  Postdoctoral Fellow very MOTIVATED

  Next Generation Detector for Protein Crystallography