Enzymatic Promiscuity

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Nelly Selem

May 2016

Approved for the Division
(Mathematics)

Francisco Barona Gomez

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

Table of Contents

Introduction **1**

Chapter 1: EvoMining **3**
1.1 Introduction 3
1.2 Gen families expansions on genomes 3
1.2.1 Pangenomes 3
1.3 EvoMining 3
1.4 Pangenome 3
1.4.1 Archaeas Results 5
1.4.2 Otras estrategias para los clusters Argon context Idea 5
1.4.3 Inline code 6
1.5 CORASON: Other genome Mining tools context-based 6
1.5.1 CORASoN 6
1.6 Loading and exploring data 7
1.7 Additional resources 10

Chapter 2: PriA Family **11**
2.1 Math 11

Chapter 3: `{r max_props} #flights %>% dplyr::filter(arr_delay == 1539, # carrier_name == "American Airlines Inc.") %>% # dplyr::select(-c(month, day, carrier, dest_name, hour, minute, carrier_name, arr_delay)) #` **19**

Chapter 4: `{r march3plot, fig.height = 3, fig.width = 6} #flights %>% dplyr::filter(month == 3, day == 3) %>% # ggplot(aes(x = dep_time, y = arr_delay)) +geom_point() #` **21**
4.1 Chemistry 101: Symbols 21
4.1.1 Typesetting reactions 23
4.1.2 Other examples of reactions 23
4.2 Physics 23
4.3 Biology 23

Chapter 5: Archaea EvoMining Results **25**
5.1 Tables 26

5.1.1 Expansions BoxPlot by metabolic family 27
5.1.2 Expansions BoxPlot by metabolic family by phylum 28
5.2 Central pathway expansions 39
5.3 Genome Size correlations 40
5.3.1 Correlation between genome size and AntiSMASH products . 40
5.3.2 Correlation between genome size and Central pathway expansions 42
5.4 Natural products 46
5.4.1 Natural products recruitments from EvoMining heatplot . . . 46
5.5 Archaeas AntiSMASH 48
5.5.1 AntisSMASH vs Central Expansions 50
5.6 Selected trees from EvoMining 53
5.7 54
5.8 Bibliographies 54
5.9 Anything else? 55

Chapter 6: Actinobacteria EvoMining Results 57
6.1 Tables 57
6.1.1 Expansions BoxPlot by metabolic family 58
6.2 Central pathway expansions 60
6.3 Genome Size correlations 64
6.3.1 Correlation between genome size and AntiSMASH products . 64
6.3.2 Correlation between genome size and Central pathway expansions 66
6.4 Natural products 70
6.4.1 Natural products recruitments from EvoMining heatplot . . . 70
6.5 Actinos AntiSMASH 72
6.5.1 AntisSMASH vs Central Expansions 74
6.6 Selected trees from EvoMining 77

Chapter 7: Cyanobacteria EvoMining Results 79
7.1 Tables 79
7.1.1 Expansions BoxPlot by metabolic family 80
7.2 Central pathway expansions 82
7.3 Genome Size correlations 83
7.3.1 Correlation between genome size and AntiSMASH products . 83
7.3.2 Correlation between genome size and Central pathway expansions 85
7.4 Natural products 89
7.4.1 Natural products recruitments from EvoMining heatplot . . . 89
7.5 Cyanobacterias AntiSMASH 91
7.5.1 AntisSMASH vs Central Expansions 93
7.6 Selected trees from EvoMining 96

Conclusion 101

Appendix A: The First Appendix 103

Appendix B: The Second Appendix, for Fun **105**

References **107**

List of Tables

1.1 Maximum Delays by Airline 9

2.1 Enzymes docking 16

5.1 Families on Archaeabacteria 26

6.1 Correlation of Inheritance Factors for Parents and Child 57

7.1 Families on Cyanobacteria 79

List of Figures

4.1 Heat Plot PriA Streptomyces vs other subtrates 22
4.2 Heat Plot TrpF Streptomyces vs other subtrates 22
4.3 Combustion of glucose 23

5.1 Expansions Boxplot 27
5.2 Archaeas Heatplot 39
5.3 Correlation between Archaeas genome size and antismash Natural products detection colored by Order 40
5.4 Correlation between Archaeas genome size and antismash Natural products detection grided by Order 41
5.5 Correlation between Archaeas genome size and central pathway expansions 42
5.6 Correlation between Archaeas genome size and central pathway expansions grided by order 43
5.7 Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family 44
5.8 Archaeas Recruitmens on central families coloured by kingdom 46
5.9 Archaeas Recruitmens on central families coloured by taxonomy . . . 47
5.10 Archaeas Diversity 48
5.11 Archaeas Smash Taxonomical Diversity 49
5.12 Correlation between Archaeas central pathway expansions and antismash Natural products detection 50
5.13 Correlation between Archaeas central pathway expnasions and antismash Natural products detection 51
5.14 Archaeas Natural products by family 52
5.15 Phosphoribosyl isomerase A EvoMiningtree 53
5.16 Phosphoribosyl isomerase other EvoMiningtree 53
5.17 Phosphoribosyl anthranilate isomerase EvoMiningtree 53

6.1 Expansions Boxplot 59
6.2 Actinobacterial Heatplot 61
6.3 Streptomyces Genomes expansions on PGA Aminoacids HeatPlot . . 63
6.4 Correlation between Actinos genome size and antismash Natural products detection colored by Order 64
6.5 Correlation between Actinos genome size and antismash Natural products detection grided by Order 65

6.6 Correlation between Actinos genome size and central pathway expansions . . . 66
6.7 Correlation between Actinos genome size and central pathway expansions grided by order . . . 67
6.8 Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family . . . 68
6.9 Actinos Recruitmens on central families coloured by kingdom . . . 70
6.10 Actinos Recruitmens on central families coloured by taxonomy . . . 71
6.11 Actinos Diversity . . . 72
6.12 Actinos Smash Taxonomical Diversity . . . 73
6.13 Correlation between Actinos central pathway expansions and antismash Natural products detection . . . 74
6.14 Correlation between Actinos central pathway expnasions and antismash Natural products detection . . . 75
6.15 Actinos Natural products by family . . . 76
6.16 Enolase EvoMiningtree . . . 77
6.17 Phosphoribosyl isomerase EvoMiningtree . . . 77
6.18 Phosphoribosyl isomerase A EvoMiningtree . . . 77
6.19 phosphoshikimate carboxyvinyltransferase EvoMiningtree . . . 78

7.1 Expansions Boxplot . . . 81
7.2 Cyanobacterial Heatplot . . . 82
7.3 Correlation between genome size and antismash Natural products detection colored by Order . . . 83
7.4 Correlation between genome size and antismash Natural products detection grided by Order . . . 84
7.5 Correlation between genome size and central pathway expansions . . . 85
7.6 Correlation between genome size and central pathway expansions grided by order . . . 86
7.7 Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family . . . 87
7.8 Recruitmens on central families coloured by kingdom . . . 89
7.9 Recruitmens on central families coloured by taxonomy . . . 90
7.10 Diversity . . . 91
7.11 Smash . . . 92
7.12 Correlation between central pathway axpnasions and antismash Natural products detection . . . 93
7.13 Correlation between central pathway axpnasions and antismash Natural products detection . . . 94
7.14 Natural products by family . . . 95
7.15 Phosphoribosyl isomerase EvoMiningtree . . . 96
7.16 Phosphoglycerate dehydrogenase EvoMiningtree . . . 96
7.17 Phosphoserine aminotransferase EvoMiningtree . . . 96
7.18 Triosephosphate isomerase EvoMiningtree . . . 97
7.19 glyceraldehyde3phosphate dehydrogenase EvoMiningtree . . . 97

7.20 phosphoglycerate kinase EvoMiningtree 97
7.21 phosphoglycerate mutaseEvoMiningtree 98
7.22 enolase EvoMiningtree 98
7.23 Pyruvate kinase EvoMiningtree 98
7.24 Aspartate transaminase EvoMiningtree 99
7.25 Asparagine synthase EvoMiningtree 99
7.26 Aspartate kinase EvoMiningtree 99
7.27 Aspartate semialdehyde dehydrogenase EvoMiningtree 100
7.28 Homoserine dehydrogenase EvoMiningtree 100

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Welcome to the *R Markdown* thesis template. This template is based on (and in many places copied directly from) the LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

Chapter 1

EvoMining

1.1 Introduction

Enzyme promiscuity on metabolic families, can be looked on enzymes that are over a divergent process.

1.2 Gen families expansions on genomes

1.2.1 Pangenomes

Expansions are located on pangenome, Tools to analyse pangenome BPgA

1.3 EvoMining

EvoMining looks expansions on prokariotic pangenome.
Biological idea.

EvoMining was available as a consult website with 230 members of the Actinobacteria phylum as genomic data base, 226 unclassified nBGCs, and not interchangable central database 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. (Cruz-Morales et al., 2016) EvoMining was proved on Actinobacteria Arseno-lipids

1.4 Pangenome

The sequenced genome of an individull in some species is just a partil print of the species geneticll repertoire Individualls can gain and loss genes.
(Koonin, 2015) Pangenome is the total sequenced gene pool in a taxonomically related group. Supergenome all the possible extant genes. About 10 times genomes. there are open, closed pangenomes.Most genomes has a core a shell and a unique genes.
Gene history its a tree history
HGT doubles mutation rate on prokarites. (Koonin, 2015) Maybe HGT is an selected

feature, if is the case, so could be np production.
Some archaeas has open pangenome. (Halachev, Loman, & Pallen, 2011) Shell trees converge to core trees (Narechania et al., 2012) ### Implementation
In order to expand **EvoMining** from a consulting website to a user-interactive genomic tool, code was developed on perl and packaged as a docker distribution. Docker is a software containerization platform that guarantee reptibilty and easy instalation.

EvoMining wiki Dockerhub nselem Docker
CORASON myRAST docker Orthocores docker *Archaea*, *Actinobacteria*, *Cyanobacteria* were used as genome DB, MIBiG was used as Natural Product DB and different Central Pathways were used.

EvoMining needs. Functional genomics uses genomes in the study of gene and protein function. 1. Genomes DB 2. Natural Products DB 3. Central Pathways DB

Genome DB

RAST annotation of genomes was done.

Phylogeny

To capture differences on genomes we sort them phylogenetically. Phylogenies can be constructed using different paradigms as Parsimony, Maximum Likelihood, and Bayesian inference. Short descriptions of the main phylogeny methods are included below. Why is a tree useful {Book reference} why trees are useful for?
* Distance methods
* Parsimony * Maximum Likelihood * Mr bayes

General Trees
Actinobacteria Tree, ArchaeaTree, CyanobacteriaTree.

It's easy to create a list. It can be unordered like

To create a sublist, just indent the values a bit (at least four spaces or a tab). (Here's one case where indentation is key!)

1. Item 1

2. Item 2
3. Item 3
    - Item 3a
    - Item 3b

Central DB

We chose central pathways from (Barona-Gómez, Cruz-Morales, & Noda-García, 2012) * BBH Best Bidirectional Hits with studied enzymes from Central Actinobacterial pathways were selected.

- By abundance
- By expansions on genomes

[largefiles,`https://help.github.com/articles/installing-git-large-file-storage/`]

Natural Products DB

Natural products was improved from previous version

1.4.1 Archaeas Results

Archaea is a kingdom of recent discovery were not many natural products has been known. On Actinobacteria, evoMining has proved its value to find new kinds of natural products. The clue to this discovery was that Actinobacteria has genomic expanssions. Now Archaea has genomic expansions, even more has central pathways genomic expansions. Are this expansions derived from a genomic duplication?
Has Archaea natural products detected by antismash, and if not, where are this NP's or may Archaea doesn't have NP's.

applying EvoMining to Archaea

1.4.2 Otras estrategias para los clusters Argon context Idea

Argon When you click the **Knit** button above a document will be generated that includes both content as well as the output of any embedded **R** code chunks within the document. You can embed an **R** code chunk like this (`cars` is a built-in **R** dataset):

```r
summary(cars)
```

```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

1.4.3 Inline code

If you'd like to put the results of your analysis directly into your discussion, add inline code like this:

> The `cos` of 2π is 1.

Another example would be the direct calculation of the standard deviation:

> The standard deviation of `speed` in `cars` is 5.2876444.

One last neat feature is the use of the `ifelse` conditional statement which can be used to output text depending on the result of an **R** calculation:

> The standard deviation is less than 6.

Note the use of `>` here, which signifies a quotation environment that will be indented.

As you see with `$2 \pi$` above, mathematics can be added by surrounding the mathematical text with dollar signs. More examples of this are in [Mathematics and Science] if you uncomment the code in Math.

1.5 CORASON: Other genome Mining tools context-based

1.5.1 CORASoN

You can also embed plots. For example, here is a way to use the base **R** graphics package to produce a plot using the built-in `pressure` dataset:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the **R** code that generated the plot. There are plenty of other ways to add chunk options. More information is available at `http://yihui.name/knitr/options/`.

Another useful chunk option is the setting of `cache = TRUE` as you see here. If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. Later in this file, you'll see a way to reference plots created in **R** or external figures.

1.6 Loading and exploring data

Included in this template is a file called `flights.csv`. This file includes a subset of the larger dataset of information about all flights that departed from Seattle and Portland in 2014. More information about this dataset and its **R** package is available at `http://github.com/ismayc/pnwflights14`. This subset includes only Portland flights and only rows that were complete with no missing values. Merges were also done with the `airports` and `airlines` data sets in the `pnwflights14` package to get more descriptive airport and airline names.

We can load in this data set using the following command:

```
flights <- read.csv("data/flights.csv")
```

The data is now stored in the data frame called `flights` in **R**. To get a better feel for the variables included in this dataset we can use a variety of functions. Here we can see the dimensions (rows by columns) and also the names of the columns.

```
dim(flights)
```

```
[1] 52808    16
```

```
names(flights)
```

```
 [1] "month"        "day"          "dep_time"     "dep_delay"
 [5] "arr_time"     "arr_delay"    "carrier"      "tailnum"
 [9] "flight"       "dest"         "air_time"     "distance"
[13] "hour"         "minute"       "carrier_name" "dest_name"
```

Another good idea is to take a look at the dataset in table form. With this dataset having more than 50,000 rows, we won't explicitly show the results of the command here. I recommend you enter the command into the Console ***after*** you have run the **R** chunks above to load the data into **R**.

```
View(flights)
```

While not required, it is highly recommended you use the `dplyr` package to manipulate and summarize your data set as needed. It uses a syntax that is easy to understand using chaining operations. Below I've created a few examples of using `dplyr` to get information about the Portland flights in 2014. You will also see the use of the `ggplot2` package, which produces beautiful, high-quality academic visuals.

We begin by checking to ensure that needed packages are installed and then we load them into our current working environment:

```
# List of packages required for this analysis
pkg <- c("dplyr", "ggplot2", "knitr", "devtools")
# Check if packages are not installed and assign the
# names of the packages not installed to the variable new.pkg
new.pkg <- pkg[!(pkg %in% installed.packages())]
# If there are any packages in the list that aren't installed,
# install them
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")
# Load packages
library(dplyr)
library(ggplot2)
library(knitr)
```

The example we show here does the following:

- Selects only the `carrier_name` and `arr_delay` from the `flights` dataset and then assigns this subset to a new variable called `flights2`.
- Using `flights2`, we determine the largest arrival delay for each of the carriers.

```
flights2 <- flights %>% dplyr::select(carrier_name, arr_delay)
max_delays <- flights2 %>% group_by(carrier_name) %>%
  summarize(max_arr_delay = max(arr_delay, na.rm = TRUE))
```

We next introduce a useful function in the `knitr` package for making nice tables in *R Markdown* called `kable`. It produces the LaTeX code required to make the table and is much easier to use than manually entering values into a table by copying and pasting values into Excel or LaTeX. This again goes to show how nice reproducible documents can be! There is no need to copy-and-paste values to create a table. (Note the use of `results = "asis"` here which will produce the table instead of the code to create the table. You'll learn more about the `\\label` later.) The `caption.short` argument is used to include a shorter version of the title to appear in the List of Tables at the beginning of the document.

```
kable(max_delays, col.names = c("Airline", "Max Arrival Delay"),
      caption = "Maximum Delays by Airline \\label{tab:max_delay}",
      caption.short = "Max Delays by Airline")
```

Table 1.1: Maximum Delays by Airline

Airline	Max Arrival Delay
Alaska Airlines Inc.	338
American Airlines Inc.	1539
Delta Air Lines Inc.	651
Frontier Airlines Inc.	575
Hawaiian Airlines Inc.	407
JetBlue Airways	273
SkyWest Airlines Inc.	421
Southwest Airlines Co.	694
United Air Lines Inc.	472
US Airways Inc.	347
Virgin America	366

We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
flights %>% dplyr::filter(arr_delay == 1539,
                    carrier_name == "American Airlines Inc.") %>%
  dplyr::select(-c(month, day, carrier, dest_name, hour,
            minute, carrier_name, arr_delay))
```

```
  dep_time dep_delay arr_time tailnum flight dest air_time distance
1     1403      1553     1934  N595AA   1568  DFW      182     1616
```

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

```
flights %>% dplyr::filter(month == 3, day == 3) %>%
  ggplot(aes(x = dep_time, y = arr_delay)) +
  geom_point()
```

1.7 Additional resources

- *Markdown* Cheatsheet - `https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet`
- *R Markdown* Reference Guide - `https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf`
- Introduction to `dplyr` - `https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html`
- `ggplot2` Documentation - `http://docs.ggplot2.org/current/`

Chapter 2

PriA Family

- Julian simulation
- Who has TrpF

```
# List of packages required for this analysis
pkg <- c("dplyr", "ggplot2", "knitr", "devtools"," reshape","RColorBrewer")
# Check if packages are not installed and assign the
# names of the packages not installed to the variable new.pkg
new.pkg <- pkg[!(pkg %in% installed.packages())]
# If there are any packages in the list that aren't installed,
# install them
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")
```

```
Warning: package ' reshape' is not available (for R version 3.3.2)
```

```
# Load packages
library(dplyr)
library(plyr)
library(reshape )
library(ggplot2)
library(knitr)
library(RColorBrewer)
hm.palette <- colorRampPalette(rev(brewer.pal(11, 'Spectral')), space='Lab')
```

2.1 Math

Docking simulation were calculated for Streptomyces enzymes
Procedures can be found at Docking Protocols

```r
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

locale:
 [1] LC_CTYPE=es_MX.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=es_MX.UTF-8        LC_COLLATE=es_MX.UTF-8
 [5] LC_MONETARY=es_MX.UTF-8    LC_MESSAGES=es_MX.UTF-8
 [7] LC_PAPER=es_MX.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_MX.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] RColorBrewer_1.1-2 reshape_0.8.6      plyr_1.8.4
[4] knitr_1.15.1       ggplot2_2.2.1      dplyr_0.5.0
[7] ape_4.0            reedtemplates_0.1  devtools_1.12.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9      magrittr_1.5     munsell_0.4.3    colorspace_1.3-2
 [5] lattice_0.20-34  R6_2.2.0         highr_0.6        stringr_1.1.0
 [9] tools_3.3.2      parallel_3.3.2   grid_3.3.2       nlme_3.1-129
[13] gtable_0.2.0     DBI_0.5-1        withr_1.0.2      htmltools_0.3.5
[17] lazyeval_0.2.0   yaml_2.1.14      rprojroot_1.2    digest_0.6.12
[21] assertthat_0.1   tibble_1.2       memoise_1.0.0    evaluate_0.10
[25] rmarkdown_1.3    labeling_0.3     stringi_1.1.2    scales_0.4.1
[29] backports_1.0.5
```

```r
library(genstats)
vignette(package="genstats")
library(devtools)
library(Biobase)
```

```
Loading required package: BiocGenerics

Loading required package: parallel


Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:parallel':

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB

The following objects are masked from 'package:dplyr':

    combine, intersect, setdiff, union

The following objects are masked from 'package:stats':

    IQR, mad, xtabs

The following objects are masked from 'package:base':

    anyDuplicated, append, as.data.frame, cbind, colnames,
    do.call, duplicated, eval, evalq, Filter, Find, get, grep,
    grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, setdiff,
    sort, table, tapply, union, unique, unsplit, which, which.max,
    which.min

Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```r
# phenoData:
tmp <- read.csv("chapter2/ProteinData", row.names = 1,header=TRUE,sep="\t")
pdata <- AnnotatedDataFrame(tmp)

  # featureData:
tmp <- read.csv("chapter2/Substrate.data", row.names = 1,sep="\t")
fdata <- AnnotatedDataFrame(tmp)

# expression data:
tmp <- read.table("chapter2/EnsymeVsSubstrate.data",row.names = 1,header=TRUE,sep=
m <- as.matrix(tmp)
#dim(m)
#class(m)
#colnames(m)
#rownames(m)
```

```
## Names should not start on numbers never
## create ExpressionSet object:
eset <- new("ExpressionSet", exprs = m, phenoData = pdata, featureData = fdata)

#pData(eset)
#fData(eset)
#vignette("01_06_three-tables")
#con=url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bodymap_eset.RData"
#load(file=con)
#close(con)
#bm = bodymap.eset
#bm
#exp_data = exprs(bm)
#dim(exp_data)
#head(exp_data,n=5)
#pheno_data = pData(bm)
#dim(pheno_data)
#head(pheno_data)
#feature_data = fData(bm)
#dim(fData(bodymap.eset))
#fData(bodymap.eset)[1:10,,1]
#devtools::session_info()
```

```
docking <- read.csv("chapter2/Heat.data", header=TRUE, sep="\t")
docking.m <- melt(docking,id = "Enzima")
#docking.m<- ddply(docking.m, .(docking.m$variable), transform, rescale=scale(value))
## NEcesito escalar!!!


ggplot(docking.m, aes(x=docking.m$variable, y=docking.m$Enzima)) + labs(x = "Substrates'
```

```
#+scale_fill_gradient(low = "white", high = "black",na.value = "orange")
#+ scale_fill_gradientn(colours = hm.palette(100),na.value = "gray")


#ggplot(docking.m, aes(x=docking.m$variable, y=docking.m$Enzima))+ geom_raster(a
#  theme(text = element_text(size=8), axis.text.x = element_text(angle = 90, hju
#  coord_equal()+scale_fill_gradient(low = "white", high = "black",na.value = "o
```

We next introduce a useful function in the `knitr` package for making nice tables in *R Markdown* called `kable`. It produces the LaTeX code required to make the table and is much easier to use than manually entering values into a table by copying and pasting values into Excel or LaTeX. This again goes to show how nice reproducible documents can be! There is no need to copy-and-paste values to create a table. (Note the use of `results = "asis"` here which will produce the table instead of the code to create the table. You'll learn more about the `\\label` later.) The `caption.short` argument is used to include a shorter version of the title to appear in the List of Tables at the beginning of the document.

```
kable(docking,  caption = "Enzymes docking \\label{tab:docking}",caption.short = "
```

Ta

Enzima	dte6_open	dte13_open	dte6_closed	dte13_closed	C04376	C03838	C04640
5AHE_Senterica	-9.1	-5.4	-6.4	-5.2	-8	-7.3	-7.4
Coli_K12	-9.7	-9.7	-9.2	-6.1	-7.2	-6.6	-6.8
Acidimicrobium	-5.7	-5.8	-6	-5.7	-6.7	-6.2	-5.8
JOAQ01	-7.4	-7.3	-7.5	-7.1	-6.5	-6.2	-6.5
2VEP	-7.5	-7.5	-7.9	-7	-7	-6.2	-6.5
2X30	-8.1	-7.4	-7.6	-6.9	-6.7	-6.8	-7.1
1VZW	na	na	na	na	na	na	na
JOFS01	-7.2	-6.7	-6.8	-6.5	-6.2	-6.6	-5.9
BAVY01	-7.4	-7.2	-7	-6.5	-7.3	-6.4	-7
JOCU01	-7.1	-7.6	-7.2	-7.3	-7.2	-6.8	-6.9
ABYA01	-9.2	-8.7	-6.7	-6.7	-7.4	-7.7	-7.2
JNXI01	-6.6	-7.3	-7	-7.1	-7.1	-7.1	-6.8
ABYC01	-6.3	-7.1	-7.4	-6.8	-7.7	-6.9	-6.6
JOFG01	-7.3	-8.8	-7.5	-6.7	-7	-6.5	-6.5
JNZV01	-8.9	-6.6	-6.7	-6.8	-7.2	-7.2	-6.2
AJUO01	-6.1	-6.8	-6.9	-6.5	-6.7	-6.3	-6.7
JNXH01	-9	-6.9	-7.1	-6.8	-6.3	-7.2	-6.6
JOFC01	-6.6	-8.2	-7.1	-6.4	-7.1	-7.4	-7.1
JNWL01	-7.4	-6.7	-7.4	-7.3	-7.7	-6.5	-6.8
AJSZ01	-6.9	-7.5	-7.9	-7.8	-7.5	-7.3	-7.7
10712	-8.3	-7.6	-7.5	-6.8	-6.4	-7	-6.2
JSFP01	-7.8	-7.2	-7.6	-7.3	-6.2	-6.5	-6.9
JJNO01	-7.4	-6.8	-7.1	-7.2	-7.2	-7.4	-7.8
JQJU01	-7.6	-7.1	-7.5	-7.4	-6.2	-6.5	-6.9
JOHB01	-7.5	-6.7	-6.8	-6.6	-6.8	-7	-6.7
JOBF01	-6.3	-6.8	-7	-6.3	-6.9	-6.9	-6.6
JNAD01	-7.3	-7	-6.8	-6.6	-6.5	-7.3	-6.6
ARLC01	-7.5	-7	-6.9	-6.6	-6.6	-6.7	-6.9
JODL01	-8	-7.3	-6.9	-6.6	-7.2	-6.7	-6.1
ADGD01	-6.4	-7.7	-7.4	-7.2	-7.6	-7.5	-7.2
ARTP01	-9.6	-10.9	-8.9	-7.1	-6	-6.2	-6.7
AEJC01	-6.6	-6.6	-7.3	-6.9	-6.5	-6.3	-6.5
ABJJ02	-5.9	-8.2	-7.1	-6.6	-6.8	-7	-7.6
4U28	na	na	na	na	na	na	na
4TX9	na	na	na	na	na	na	na
JOBU01	-8	-7.1	-6.7	-6.9	-6.6	-6.8	-6.7
JNZG01	-7.8	-7.2	-7.6	-7.2	-7.4	-6.9	-7.4
JNZY01	-6.4	-7	-7.1	-6.7	-7.7	-6.3	-6.2
JNZH01	-7.5	-6.9	-6.8	-6.8	-6.5	-6.6	-6.5
ACEW01	-7.4	-6.9	-7.3	-6.8	-7.3	-6.6	-7.5
ATCJ01	-6.5	-6.9	-7.2	-7.1	-6.5	-6.9	-6.4

Enzima	dte6_open	dte13_open	dte6_closed	dte13_closed	C04376	C03838
4X9S	na	na	na	na	na	na
4W9T	na	na	na	na	na	na
AWQW01	-6.2	-6.8	-7.4	-6.5	-7.8	-6.8
JOFK01	-7.6	-6.7	-7.1	-7.2	-7.1	-6.5
JODS01	-6.8	-7.3	-6.9	-6.7	-8	-7.9
TC1	-8.2	-8.5	-8.1	-7.9	-5.7	-5.5
4WD0_TC1	na	na	na	na	na	na
Auro 4X2R	-9.9	-9.1	-9.8	-8.1	-7.4	-7.1
Acardi_b2	-10.6	-9.3	-9.3	-7.2	-7.2	-6.8
JNZF01	-6.9	-6.8	-7.5	-7.1	-7.8	-7.3
2Y88	-10	-7.8	-8.9	-5.4	-8.6	-7.3
2Y89	-8.7	-8.6	-9.6	-9.4	-6.4	-5.9
2Y85	-8.2	-7.9	-9.2	-7.4	-7.6	-7.5
3ZS4	-10	-10.5	-11.4	-6.4	-8.2	-7.2
diphtheriae	-9.2	-7.8	-10.9	-7.2	-7.5	-7.6
jeikeium	-7.5	-6.9	-7.2	-6.5	-8.4	-8
JSFP01_2	-7.4	-8	-7.6	-6.4	-5.2	-5.4
TrpF Jonesia 4WUI	-8.3	-8.8	-8.2	-6.8	-6.2	-6.1
trachomatis	-8.2	-8	-7.4	-7.2	-6.1	-5.5
mg1	-7.2	-8.8	-8.2	-7.3	-6.2	-5.7
Aodo_b12	-8.5	-8.6	-8.8	-7.3	-7.1	-7.1

We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

Chapter 3

```
{r max_props} #flights %>%
dplyr::filter(arr_delay == 1539,
#                    carrier_name
== "American Airlines Inc.") %>%
#  dplyr::select(-c(month, day,
carrier, dest_name, hour, minute,
carrier_name, arr_delay)) #
```

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

Chapter 4

{r march3plot, fig.height = 3, fig.width = 6} #flights %>% dplyr::filter(month == 3, day == 3) %>% # ggplot(aes(x = dep_time, y = arr_delay)) +geom_point()

Genome size vs Total antismash cluster coloured by order

Docker simulation were calculated for Streptomyces enzymes

Genome size vs Total antismash cluster coloured by order

TEX is the best way to typeset mathematics. Donald Knuth designed TEX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read LATEX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section which has been commented out. If you're not going to use math, skip over or delete this next commented section.

4.1 Chemistry 101: Symbols

Chemical formulas will look best if they are not italicized. Get around math mode's automatic italicizing in LATEX by using the argument `$\mathrm{formula here}$`, with your formula inside the curly brackets. (Notice the use of the backticks here which enclose text that acts as code.)

So, $Fe_2^{2+}Cr_2O_4$ is written `$\mathrm{Fe_2^{2+}Cr_2O_4}$`.

Exponent or Superscript: O^-

Subscript: CH_4

To stack numbers or letters as in Fe_2^{2+}, the subscript is defined first, and then the

Figure 4.1: Heat Plot PriA Streptomyces vs other subtrates

Figure 4.2: Heat Plot TrpF Streptomyces vs other subtrates

superscript is defined.
Angstrom: Å
Bullet: $CuCl \bullet 7H_2O$
Double Dagger: ‡
Delta: Δ
Reaction Arrows: $\longrightarrow$ or $\xrightarrow{solution}$
Resonance Arrows: $\leftrightarrow$
Reversible Reaction Arrows: $\rightleftharpoons$ or $\xrightleftharpoons{solution}$ (the latter requires the `chemarr` LaTeX package which is automatically loaded in this template)

4.1.1 Typesetting reactions

You may wish to put your reaction in a figure environment, which means that LaTeX will place the reaction where it fits and you can have a figure caption. You'll see further description of this **R** `label` function in . (Note the use of the double backslash here as well as the `echo = FALSE` which hides the code from the output.)

$$C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O$$

Figure 4.3: Combustion of glucose

4.1.2 Other examples of reactions

$NH_4Cl_{(s)} \rightleftharpoons NH_{3(g)} + HCl_{(g)}$
$MeCH_2Br + Mg \xrightarrow[below]{above} MeCH_2 \bullet Mg \bullet Br$

4.2 Physics

Many of the symbols you will need can be found on the math page `http://web.reed.edu/cis/help/latex/math.html` and the Comprehensive LaTeX Symbol Guide (`http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf`).

4.3 Biology

You will probably find the resources at `http://www.lecb.ncifcrf.gov/~toms/latex.html` helpful, particularly the links to bsts for various journals. You may also be interested in TeXShade for nucleotide typesetting (`http://homepages.uni-tuebingen.de/beitz/txe.html`). Be sure to read the proceeding chapter on graphics and tables.

Chapter 5

Archaea EvoMining Results

During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eucarya in an exciting first great application of 16S phylogeny(C R Woese & Fox, 1977,Carl R. Woese & Gupta (1981)) . Main differences between this kingdoms are that Archaeal DNA is not arranged in a nucleus as in Eucarya and Archaeal celular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be higlhy valuable to biotechnology industry for their great stability due to extreme temperature, PH and salt content conditions on Archeal habitats. Despite no Archaeal Natural products biosynthetic gene clusters (BGC's) has been reported on MiBIG, Archaea do have BGC's, some of them seems to be acquired by horizontal gene transfer (HGT) like methano nrps {search reference}. Other Archeal natural products known are archaeosins, Diketopiperazines, Acyl Homoserine Lactones, Exopolysaccharides, Carotenoids, Biosurfactants, Phenazines and Organic Solutes but this knowledge is not comparable to Bacterial BGC's knowledge(Charlesworth & Burns, 2015).

Natural products biosynthetic gene clusters search is actually performed using either *high-confidence/low-novelty or low-confidence/high-novelty* bioinformatic approaches (Medema & Fischbach, 2015). High confidence methods compares query sequences with previously known BGC's such as nrps or PKS, examples of this algorithms are antiSMASH and clusterfinder (**???**?). EvoMining searches on expansions from central metabolic pathways enzyme families, it has been classified as low confidence/high novelty method. EvoMining has proved useful on Actinobacteria phylum where its use lead to Arseno-compounds discovery (Cruz-Morales et al., 2016). Also on Actinobacteria antiSMASH analysis on 1245 genomes found 774 different classes of natural products, the same analysis on 876 Archaeal genomes, a full kingdom, identifies only 35 BGC's classes. So either Archaea does not have natural products BGC's or this are not yet known. Next paragraph deals with a possible approach about how natural products BGC's can be find.

Archaea resembled Bacteria in that Archaea uses horizontal gene transfer as a genic interchange mecanism, Archaeal genomes contains operons (Howland, 2000) and in general there is introns absence{Reference to Computational Methods for Understanding Bacterial and Archaeal Genomes}. Archaeas do have introns, but they are mainly located on genes that encodes ribosomal and transfer RNA (Howland,

2000). General lack of introns allows automatic genome annotation, operons gene organization permits functional inference to a certain degree and HGT contribute to expansions on Archaeal genomes. Some phylum on Archaea has an open pangenome, and as we will show on this chapter some Archaea has central pathway expansions. Enzyme families from central pathways expansions, open pangenome and operon organization made EvoMining succesful on Actinobacteria, this lead us to think that evoMining is suitable to analize Archaeal genomes, even more since EvoMining is a method oriented to use evolution and its not entirelyy based on previous knowledge of BGC's sequences if evolutionary logic behave on Archaea as on bacteria, new BGC's classes may be be found on Archaea.

EvoMining is a trade off between conserved known central metabolic function and enough expansions divergence on sequence and on clusters to divergence

5.1 Tables

Table 5.1: Families on Archaeabacteria

Factors	Correlation between Parents & Child
GenomeDB	876
Phylum	12
Order	23

First lets investigate if Archaea has expansions on families within central metabolic routes. Since main metabolic pathways are shared between Bacteria and Archaea makes sense to assemble Archeal EvoMining central database by using orthologous from Actinobacteria evoMining central pathways.

5.1.1 Expansions BoxPlot by metabolic family

```
label(path = "chapter3/expansion_plotArchaeas.pdf", caption = "Expansions Boxplot"
```

Figure 5.1: Expansions Boxplot

Here is a reference to the expansion boxplot: Figure 5.1.

5.1.2 Expansions BoxPlot by metabolic family by phylum

```
#+ geom_jitter()
#aes(fill = factor(vs))

ArchaeasTotalBP.m<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x="RastId",by.y="RastId") ## w
ArchaeasHeatPlotBP.m <- melt(ArchaeasTotalBP.m,id =c("RastId","Name","SuperPhylum","Phyl
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expecte
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expecte

## Each metabolic pathway se parte por phylum coloreado por order

#3PGA_AMINOACIDS
#Glycolysis
#OXALACETATE_AMINOACIDS
#R5P_AMINOACIDS
#TCA
#E4P_AMINO_ACIDS
#PYR_THR_AA

## Genome size
ggplot(ArchaeasHeatPlotBP.m, aes(x=ArchaeasHeatPlotBP.m$Phylum, y=ArchaeasHeatPlotBP.m$S
```

```
#+ geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))


## Halobacteria
MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Phylum=="Euryarchaeota")
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Order, y=MetFam_BP.m$Size))+ geom_boxplot()
```

```
#MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Family=="Methanosarcinaceae")
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
#+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size


  #geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))# + facet_grid(. ~ Phylum)+theme


## Metabolic Pathways
MetFam=subset(ArchaeasCentral,Pathway=="PPP")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="3PGA_AMINOACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="Glycolysis")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="OXALACETATE_AMINOACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="R5P_AMINOACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

#
MetFam=subset(ArchaeasCentral,Pathway=="TCA")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="E4P_AMINO_ACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))


MetFam=subset(ArchaeasCentral,Pathway=="PYR_THR_AA")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

#ggsave("chapter3/expansion_plotArchaeas.pdf", plot = expansion_plotArchaea,height = 8
```

5.2 Central pathway expansions

Heat plot of central pathways expansions, Needs to be phylogenetically sorted.

Figure 5.2: Archaeas Heatplot

Here is a reference to the HeatPlot: Figure 5.2.

5.3 Genome Size correlations

5.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

Figure 5.3: Correlation between Archaeas genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 5.3.

Genome size vs Total antismash cluster detected splitted by order

Figure 5.4: Correlation between Archaeas genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 5.4.

5.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

Figure 5.5: Correlation between Archaeas genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 5.5.

Genome size vs Total central pathway expansion grided by order

Figure 5.6: Correlation between Archaeas genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 5.6.

Correlation between genome size and each of the central pathway families. Data are coloured by metabolic family instead of coloured by taxonomical order. This treatment allows to answer how differente metabolic families grows when genome size grow.
Also I want to add form given by taxonomical order.

```
Warning: The shape palette can deal with a maximum of 6 discrete values
because more than 6 becomes difficult to discriminate; you have
24. Consider specifying shapes manually if you must have them.
```

```
Warning: Removed 64823 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

Figure 5.7: Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 5.7.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

5.4 Natural products

5.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

Figure 5.8: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 5.8.

Recruitments after central pathways expansions colourd by taxonomy

Figure 5.9: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 5.9.

5.5 Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

Figure 5.10: Archaeas Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 5.10.

Smash diversity

Figure 5.11: Archaeas Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 5.11.

5.5.1 AntisSMASH vs Central Expansions

Is it a correlation between pangenome grow and central pathways expansions?

Total central pathway expansions by genome vs Total antismash cluster detected coloured by order

Figure 5.12: Correlation between Archaeas central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 5.12.

Total central pathway expansions by genome vs Total antismash cluster detected splitted by order

Figure 5.13: Correlation between Archaeas central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 5.13.

AntisMAsh vs Expansions by taxonomic Family
Natural products colured by family

Figure 5.14: Archaeas Natural products by family

Here is a reference to the Natural products colured by family plot Figure 5.14.

5.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

Figure 5.15: Phosphoribosyl isomerase A EvoMiningtree

Figure 5.16: Phosphoribosyl isomerase other EvoMiningtree

Figure 5.17: Phosphoribosyl anthranilate isomerase EvoMiningtree

5.7

Other possible databases Archaeal signatures *set of protein-encoding genes that function uniquely within the Archaea; most signature proteins have no recognizable bacterial or eukaryal homologs* (Graham, Overbeek, Olsen, & Woese, 2000) ## Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

5.8 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (Molina & Borkovec, 1994). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

[1] footnote text

[2] Reed College (2007)

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

5.9 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

[3] Noble (2002)

Chapter 6

Actinobacteria EvoMining Results

Actinobacteria is an ancient phylum {Referencia de luis}

6.1 Tables

Table 6.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

6.1.1 Expansions BoxPlot by metabolic family

```
label(path = "chapter4/expansion_plotActinos.pdf", caption = "Expansions Boxplot",label
```

Here is a reference to the expansion boxplot: Figure 6.1.

Figure 6.1: Expansions Boxplot

6.2 Central pathway expansions

Heat plot of central pathways expansions, Needs to be phylogenetically sorted.

Here is a reference to the HeatPlot: Figure 6.2.

PPP pahtway expansions restricted to *Streptomycetaceae* family HeatPlot: Figure 6.2.

Here is a reference to the HeatPlot: Figure 6.3.

288310
288306
288302
288298
288289
288280
288278
288277
288268
288267
288266
288261
288254
288233
288231
288218
288215
288188
288178
288152
288032
288019
288012
287996
287981
287979
287965
287955
287953
287950
287948
262855
252178
252176
252172
252171
252170
252168
252167
252165
242654
242652
242564
242561
242560
242557
242552
242551
242548
242547
242546
242545
242543
242541
242539
242537
242535
242532
242530
242529
242528
242522
242521
242515
242513
242511
242507
242502
242501
242500
242499
242491
242490
242488
242486
242480
242479
242478
242476
242475
242474
242471
242470
242465
242464
242463
242452
242451
242449
242448
242445
242444
242442
242441
242440
242438
242437
242435
242434
242433
242428
242426
242425
242419
242417
242415
242412
242410
242407
242404
242402
242397
242396
242393
242391
242390
242388
242386
242385
242384
242383
242380
242378
242377
242373
242370
242365
242364
242363
242357
242352
242351
242350
242347
242344
242340
242333
242331
242325
242320
242319
242318
242312
242311
242310
242307
242305
242300
242298
242291
242287
242286
242285
242272
242271
242268
242266
242265
242263

6.3 Genome Size correlations

6.3.1 Correlation between genome size and AntiSMASH products

```
Warning: Removed 1 rows containing missing values (geom_point).

Warning: Removed 1 rows containing missing values (geom_point).
```

Genome size vs Total antismash cluster coloured by order

Figure 6.4: Correlation between Actinos genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 6.4.

Genome size vs Total antismash cluster detected splitted by order

Figure 6.5: Correlation between Actinos genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 6.5.

6.3.2 Correlation between genome size and Central pathway expansions

```
Warning: Removed 1 rows containing missing values (geom_point).

Warning: Removed 1 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by order

Figure 6.6: Correlation between Actinos genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 6.6.

Genome size vs Total central pathway expansion grided by order

Figure 6.7: Correlation between Actinos genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 6.7.

Correlation between genome size and each of the central pathway families. Data are coloured by metabolic family instead of coloured by taxonomical order. This treatment allows to answer how differente metabolic families grows when genome size grow.
Also I want to add form given by taxonomical order.

```
Warning: The shape palette can deal with a maximum of 6 discrete values
because more than 6 becomes difficult to discriminate; you have
32. Consider specifying shapes manually if you must have them.

Warning: Removed 103306 rows containing missing values (geom_point).

Warning: Removed 94 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

Figure 6.8: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 6.8.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

6.4 Natural products

6.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

Figure 6.9: Actinos Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 6.9.

Recruitments after central pathways expansions colourd by taxonomy

Figure 6.10: Actinos Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 6.10.

6.5 Actinos AntiSMASH

Taxonomical diversity on Actinosbacteria Data

Figure 6.11: Actinos Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 6.11.

Smash diversity

Figure 6.12: Actinos Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 6.12.

6.5.1 AntisSMASH vs Central Expansions

Is it a correlation between pangenome grow and central pathways expansions?

Total central pathway expansions by genome vs Total antismash cluster detected coloured by order

Figure 6.13: Correlation between Actinos central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 6.13.

Total central pathway expansions by genome vs Total antismash cluster detected splitted by order

Figure 6.14: Correlation between Actinos central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 6.14.

AntisMAsh vs Expansions by taxonomic Family
Natural products colured by family

Figure 6.15: Actinos Natural products by family

Here is a reference to the Natural products colured by family plot Figure 6.15.

6.6 Selected trees from EvoMining

Figure 6.16: Enolase EvoMiningtree

Figure 6.17: Phosphoribosyl isomerase EvoMiningtree

Figure 6.18: Phosphoribosyl isomerase A EvoMiningtree

Figure 6.19: phosphoshikimate carboxyvinyltransferase EvoMiningtree

Chapter 7

Cyanobacteria EvoMining Results

Cyanobacteria is a photosynthetic phylum that inhabits a broad range of habitats. The broad adaptive potential is on part driven by gene-family enlargment (Larsson, Nylander, & Bergman, 2011) by the analysis of 58 Cyanobacterial genomes concludes ancestor of cyanobacteria had a genome size of approx. 4.5 Mbp. Cyanobacteria produces natural products as pigments and toxins (Whitton, 2012) Example of a PriA cluster toxins(Moustafa et al., 2009)

Fossil record situates Cyanobacteria (Whitton, 2012) Molecular record and metabolic propoerties at (Battistuzzi, Feijao, & Hedges, 2004)

7.1 Tables

Table 7.1: Families on Cyanobacteria

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

7.1.1 Expansions BoxPlot by metabolic family

```
label(path = "chapter5/expansion_plotCyanos.pdf", caption = "Expansions Boxplot",label =
```

Here is a reference to the expansion boxplot: Figure 7.1.

Figure 7.1: Expansions Boxplot

7.2 Central pathway expansions

Heat plot of central pathways expansions, Needs to be phylogenetically sorted.

Figure 7.2: Cyanobacterial Heatplot

Here is a reference to the HeatPlot: Figure 7.2.

7.3 Genome Size correlations

7.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

Figure 7.3: Correlation between genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 7.3.

Genome size vs Total antismash cluster detected splitted by order

Figure 7.4: Correlation between genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 7.4.

7.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

Figure 7.5: Correlation between genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 7.5.

Genome size vs Total central pathway expansion grided by order

Figure 7.6: Correlation between genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 7.6.

Correlation between genome size and each of the central pathway families. Data are coloured by metabolic family instead of coloured by taxonomical order. This treatment allows to answer how differente metabolic families grows when genome size grow.
Also I want to add form given by taxonomical order.

```
Warning: The shape palette can deal with a maximum of 6 discrete values
because more than 6 becomes difficult to discriminate; you have
10. Consider specifying shapes manually if you must have them.
```

```
Warning: Removed 20418 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

Figure 7.7: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 7.7.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

7.4 Natural products

7.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

Figure 7.8: Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 7.8.

Recruitments after central pathways expansions colourd by taxonomy

Figure 7.9: Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 7.9.

7.5 Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

Figure 7.10: Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 7.10.

Smash diversity

Figure 7.11: Smash

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: **??**.

7.5.1 AntisSMASH vs Central Expansions

Is it a correlation between pangenome grow and central pathways expansions?

Total central pathway expansions by genome vs Total antismash cluster detected coloured by order

Figure 7.12: Correlation between central pathway axpnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 7.12.

Total central pathway expansions by genome vs Total antismash cluster detected splitted by order

Figure 7.13: Correlation between central pathway axpnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot ??.

AntisMAsh vs Expansions by taxonomic Family
Natural products colured by family

Figure 7.14: Natural products by family

Here is a reference to the Natural products colured by family plot Figure 7.14.

7.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

Figure 7.15: Phosphoribosyl isomerase EvoMiningtree

Figure 7.16: Phosphoglycerate dehydrogenase EvoMiningtree

Figure 7.17: Phosphoserine aminotransferase EvoMiningtree

Figure 7.18: Triosephosphate isomerase EvoMiningtree

Figure 7.19: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

Figure 7.20: phosphoglycerate kinase EvoMiningtree

Figure 7.21: phosphoglycerate mutaseEvoMiningtree

Figure 7.22: enolase EvoMiningtree

Figure 7.23: Pyruvate kinase EvoMiningtree

Figure 7.24: Aspartate transaminase EvoMiningtree

Figure 7.25: Asparagine synthase EvoMiningtree

Figure 7.26: Aspartate kinase EvoMiningtree

Figure 7.27: Aspartate semialdehyde dehydrogenase EvoMiningtree

Figure 7.28: Homoserine dehydrogenase EvoMiningtree

Conclusion

Idea de Rosario -ver dell cluster de saxitoxin cuantos pasos se necesitron para llegar ahi.
-A donde se iria el resultado de abrir el GMP
-Otra vez, que Actinos tienen FolE

If we don't want Conclusion to have a chapter number next to it, we can add the `{.unnumbered}` attribute. This has an unintended consequence of the sections being labeled as 3.6 for example though instead of 4.1. The LaTeX commands immediately following the Conclusion declaration get things back on track.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file:

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
}
library(reedtemplates)
```

In :

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(plyr))
    install.packages("plyr", repos = "http://cran.rstudio.com")
if(!require(dplyr))
    install.packages("dplyr", repos = "http://cran.rstudio.com")
```

```r
if(!require(ggplot2))
    install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
  }
library(reedtemplates)
flights <- read.csv("data/flights.csv")
```

Appendix B

The Second Appendix, for Fun

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Barona-Gómez, F., Cruz-Morales, P., & Noda-García, L. (2012). What can genome-scale metabolic network reconstructions do for prokaryotic systematics? *Antonie van Leeuwenhoek*, *101*(1), 35–43. http://doi.org/10.1007/s10482-011-9655-1

Battistuzzi, F. U., Feijao, A., & Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, *4*, 44. http://doi.org/10.1186/1471-2148-4-44

Charlesworth, J. C., & Burns, B. P. (2015). Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. *Archaea*, *2015*, e282035. http://doi.org/10.1155/2015/282035

Cruz-Morales, P., Kopp, J. F., Martínez-Guerrero, C., Yáñez-Guerra, L. A., Selem-Mojica, N., Ramos-Aboites, H., ... Barona-Gómez, F. (2016). Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biology and Evolution*, *8*(6), 1906–1916. http://doi.org/10.1093/gbe/evw125

Graham, D. E., Overbeek, R., Olsen, G. J., & Woese, C. R. (2000). An archaeal genomic signature. *Proceedings of the National Academy of Sciences*, *97*(7), 3304–3308. http://doi.org/10.1073/pnas.97.7.3304

Halachev, M. R., Loman, N. J., & Pallen, M. J. (2011). Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. *PLOS ONE*, *6*(12), e28388. http://doi.org/10.1371/journal.pone.0028388

Howland, J. L. (2000). *The surprising archaea: Discovering another domain of life*. New York: Oxford University.

Koonin, E. V. (2015). The Turbulent Network Dynamics of Microbial Evolution

and the Statistical Tree of Life. *Journal of Molecular Evolution*, *80*(5-6), 244–250. http://doi.org/10.1007/s00239-015-9679-7

Larsson, J., Nylander, J. A., & Bergman, B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology*, *11*, 187. http://doi.org/10.1186/1471-2148-11-187

Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature Chemical Biology*, *11*(9), 639–648. http://doi.org/10.1038/nchembio.1884

Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley.

Moustafa, A., Loram, J. E., Hackett, J. D., Anderson, D. M., Plumley, F. G., & Bhattacharya, D. (2009). Origin of Saxitoxin Biosynthetic Genes in Cyanobacteria. *PLOS ONE*, *4*(6), e5758. http://doi.org/10.1371/journal.pone.0005758

Narechania, A., Baker, R. H., Sit, R., Kolokotronis, S.-O., DeSalle, R., & Planet, P. J. (2012). Random Addition Concatenation Analysis: A Novel Approach to the Exploration of Phylogenomic Signal Reveals Strong Agreement between Core and Shell Genomic Partitions in the Cyanobacteria. *Genome Biology and Evolution*, *4*(1), 30–43. http://doi.org/10.1093/gbe/evr121

Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.

Reed College. (2007, march). LaTeX your document. Retrieved from http://web.reed.edu/cis/help/LaTeX/index.html

Whitton, B. A. (2012). *Ecology of Cyanobacteria II: Their Diversity in Space and Time*. Springer Science & Business Media.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/

Woese, C. R., & Gupta, R. (1981). Are archaebacteria merely derived /"prokaryotes/"? *Nature*, *289*(5793), 95–96. http://doi.org/10.1038/289095a0