

Codon and aminoacid usage in bacteria
(HAECKEL, 1874)

P^r Jean R. LOBRY

Work in progress. This is release : June 10, 2019
Pre- α release tag

WARNING: this is not a peer-reviewed nor a genuine published piece of work. This is just a draft, a work in progress. I'm planning to have an α -version for spring 2020, a β -version for spring 2021, and hopefully a decent version in spring 2022. In the meantime, feedback is more than welcome. Just look for the string "jean lobry" in your favorite internet search engine to get my e-mail.



Contents

1	Introduction	1
1.1	Nature of the document	1
1.2	Structure of the document	1
2	Univariate analysis of aminoacid usage	3
2.1	Loading the dataset	3
2.2	Introduction	3
2.2.1	GC content definition and properties	3
2.2.2	SUEOKA's plots	4
2.2.3	GC content as a nuisance parameter	7
2.2.4	Aminoacid frequencies under neutral conditions	8
2.2.5	Aminoacid classes with respect to GC content	10
2.3	Class 1 aminoacids	17
2.3.1	Isoleucine	17
2.3.2	Phenylalanine	18
2.3.3	Lysine	22
2.3.4	Tyrosine	25
2.3.5	Asparagine	25
2.3.6	Leucine	27
2.4	Class 2 aminoacids	30
2.4.1	Methionine	30
2.4.2	Aspartic acid	32
2.4.3	Glutamic acid	34
2.4.4	Serine	36
2.4.5	Valine	39
2.4.6	Threonine	43
2.4.7	Histidine	47
2.4.8	Glutamine	49
2.4.9	Cysteine	51
2.4.10	Tryptophane	51
2.5	Class 3 aminoacids	52
2.5.1	Arginine	52
2.5.2	Alanine	54
2.5.3	Proline	56
2.5.4	Glycine	58
2.6	Evolution of hydrolysis sensitive aminoacids with GC content	60
2.6.1	Aspartic acid and asparagine	60
2.6.2	Glutamic acid and glutamine	62

2.7	Evolution of charged aminoacids with GC content	64
2.7.1	Negatively charged aminoacid	64
2.7.2	Positively charged aminoacid	64
2.7.3	Evolution of pI with GC	65
2.8	Summary of outstanding bacterial groups	67
3	Multivariate analysis of aminoacid usage	69
3.1	Loading the dataset	69
3.2	Utilities	69
3.2.1	First factorial map orientation	69
3.3	Sanity check	69
3.3.1	Direct CA on aminoacid frequencies	69
3.3.2	BCA on codon frequencies	70
3.3.3	Comparisons	71
3.3.4	Conclusion	71
4	Univariate analysis of synonymous codon usage	73
4.1	Utilities definition	73
4.1.1	Loading the dataset	73
4.1.2	Computing codon relative frequencies	73
4.1.3	Ploting data	73
4.1.4	Generation of all figures	74
4.2	Introduction	74
4.3	Terminators	75
4.4	Odd number	76
4.5	Duet	78
4.5.1	Asparagine	78
4.5.2	Aspartic acid	78
4.5.3	Cysteine	79
4.5.4	Glutamine	79
4.5.5	Glutamic acid	80
4.5.6	Histidine	80
4.5.7	Lysine	81
4.5.8	Phenylalanine	81
4.5.9	Tyrosine	82
4.6	Quartet	83
4.6.1	Alanine	83
4.6.2	Glycine	84
4.6.3	Proline	85
4.6.4	Threonine	86
4.6.5	Valine	87
4.7	Sextet	88
4.7.1	Arginine	88
4.7.2	Leucine	89
4.7.3	Serine	90
5	Multivariate analysis of synonymous codon usage	91
5.1	Loading the dataset	91

6	Dataset compilation	93
6.1	Introduction	93
6.1.1	Purpose	93
6.1.2	Bacterial growth as function of temperature	93
6.2	Origin of data	101
6.2.1	T_{opt} data from ENGQVIST 2018	101
6.2.2	Codon usage data from LOBRY 2018	102
6.2.3	T_{opt} data from LOBRY & NECŞULEA 2006	104
6.2.4	T_{opt} data from GALTIER & LOBRY 1997	104
6.3	T_{opt} curation	105
6.3.1	Merging tables	105
6.3.2	Taxonomic filtering	105
6.3.3	Available T_{opt} before curation	106
6.3.4	T_{opt} comparison between [31] and [38, 73]	106
6.3.5	Solving important T_{opt} discrepancies ($> 5^{\circ}\text{C}$)	108
6.3.6	Collation finale des températures optimales de croissance	122
6.3.7	Manual bibliographical search for T_{opt}	123
6.4	Polishing data	129
6.4.1	Sélection des lignes et colonnes, tri et sauvegarde	129
6.4.2	GC content computation	135
6.4.3	Computing aminoacid frequencies	138
6.4.4	Computing isoelectric points	138
6.4.5	Backup	144
7	Conclusion	145
8	Annexes	147
8.1	Code for figures	147
8.2	Session information	151
	Bibliography	153

Chapter 1

Introduction

1.1 Nature of the document

1.2 Structure of the document

Chapter 2

Univariate analysis of aminoacid usage

2.1 Loading the dataset

```
load("local/tdd.Rda")
```

2.2 Introduction

2.2.1 GC content definition and properties

CONSIDER a doubled-stranded DNA genome. Pick one strand, let call it the plus-strand, and assume that its primary chemical formula is given by:

$$A_{a_+} C_{c_+} G_{g_+} T_{t_+} \quad (2.1)$$

where $(a_+, c_+, g_+, t_+) \in \mathbb{N}^4$ are the total number of the four bases in the plus-strand. For bacteria there are typically in 10^6 units. The GC content of the plus-strand, θ_+ , usually expressed in percent, is the relative frequency of bases G or C:

$$\theta_+(a_+, c_+, g_+, t_+) = 100 \times \frac{g_+ + c_+}{a_+ + c_+ + g_+ + t_+} \quad (2.2)$$

CONSIDER the complementary strand of the plus-strand, call it the minus-strand, and assume using analogous notations that its primary formula is given by:

$$A_{a_-} C_{c_-} G_{g_-} T_{t_-} \quad (2.3)$$

where $(a_-, c_-, g_-, t_-) \in \mathbb{N}^4$ are the total number of the four bases in the minus-strand. The GC content of the minus-strand is given by:

$$\theta_-(a_-, c_-, g_-, t_-) = 100 \times \frac{g_- + c_-}{a_- + c_- + g_- + t_-} \quad (2.4)$$

Now, from the structure of the doubled-stranded DNA molecule [151] it follows that the number of A in the plus-strand equals the number of T in the minus-strand, $a_+ = t_-$ (*vice versa* $a_- = t_+$) and that the number of G in the plus-strand equals the number of C in the minus-strand, $g_+ = c_-$ (*vice versa* $g_- = c_+$):

$$\begin{cases} a_+ = t_- \\ a_- = t_+ \\ g_+ = c_- \\ g_- = c_+ \end{cases} \quad (2.5)$$

The direct consequence is that the GC content is exactly the same in the two strands of a double-stranded DNA molecule, as can be seen by using the equalities in 2.6 to transform equations 2.2 and 2.4:

$$\left\{ \begin{array}{l} \theta_+(a_+, c_+, g_+, t_+) = \theta_-(a_-, c_-, g_-, t_-) \\ \Leftrightarrow \frac{g_+ + c_+}{a_+ + c_+ + g_+ + t_+} = \frac{g_- + c_-}{a_- + c_- + g_- + t_-} \\ \Leftrightarrow \frac{g_+ + c_+}{a_+ + c_+ + g_+ + t_+} = \frac{c_+ + g_+}{t_+ + g_+ + c_+ + a_+} \\ \Leftrightarrow \frac{g_+ + c_+}{a_+ + c_+ + g_+ + t_+} = \frac{g_+ + c_+}{a_+ + c_+ + g_+ + t_+} \\ \Leftrightarrow 1 = 1 \end{array} \right. \quad (2.6)$$

In bacterial genomes there is a wide variation of the GC content, ranging from ~25% to ~75% [64, 11, 137]. The amount of intragenomic variability is at contrast very small [135, 116, 138]. The within-species variability of GC content is low [16] but this is somewhat circular because the GC content is one of the genomic characteristics recommended for the description of bacterial species and genera. To give a rough idea, 5% and 10% are the common range of GC content variation found within a species and a genera, respectively. The wide inter-species variation and narrow intra-species heterogeneity of the GC content was interpreted as the result of bidirectional mutation rates between AT and GC pairs in SUEOKA's directional mutation pressure theory [137]. He was the first to state in 1962, before the emergence of neutralism, that some patterns of the genome could appear without natural selection, a paradigm switch at that time.

BECAUSE in double-stranded DNA G-C base-pairing is stronger (3 hydrogen bounds) than A-T base-pairing (2 hydrogen bounds), the GC content can be estimated easily by measuring the temperature at which the DNA melts [79]. Due to the thermostability given to high GC DNA, it was commonly believed that the GC content played a role in adaptation at high temperatures, a hypothesis that was refuted in 1997 [38] and figure 2.1 page 5 shows that it is still the case.

2.2.2 Sueoka's plots

The influence of genomic GC content on the average aminoacid composition of proteins was pioneered by SUEOKA (1961) [136]. The conclusion of the paper stated that "[t]here exist several significant correlations between DNA base composition and amino acid composition of proteins. Among 18

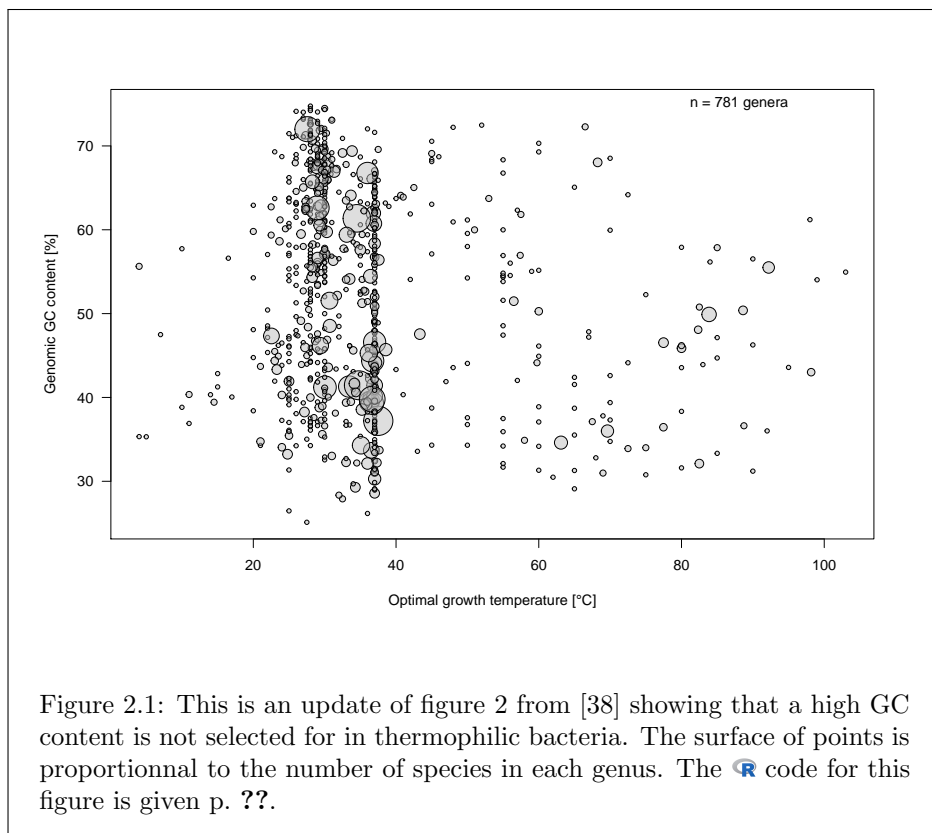


Figure 2.1: This is an update of figure 2 from [38] showing that a high GC content is not selected for in thermophilic bacteria. The surface of points is proportionnal to the number of species in each genus. The `R` code for this figure is given p. ??.

amino acid tested, alanine, arginine, glycine, and proline are positively correlated with guanine-cytosine content of DNA. Isoleucine, lysine, aspartic acid plus asparagine, glutamic acid plus glutamine, tyrosine and phenylalanine are negatively correlated. Histidine, valine, leucine, threonine, serine and possibly methionine are extremely uniform with no detectable evidence correlation. The results obtained were discussed in the relation to the coding problem¹.”

Make a \LaTeX table with table 1 from [136]

do not confuse with neutrality plot

WHAT I call a SUEOKA’s plot here is what was used to draw figure 1 in [136], that is aminoacid relative frequency as function of the GC content. It’s easy to re-create the figures because data are given in table 1 from [136]². Note that we cannot compare directly with present data for several reasons:

1. A bulk protein extract is enriched in highly expressed genes products, that is mainly ribosomal proteins in bacteria. This is not the same as giving the same weight to all protein coding genes when using complete genome data. In *Escherichia coli*, for instance, the average composition of the products of genes with a high expressivity is known to be different, even if the associated variability is less than the one due to the opposition between integral membrane proteins and cytoplasmic proteins [72].
2. The hydrolysis of the peptidic bounds will also target the amide bounds in the side chain of Asn and Gln yielding Asp and Glu, respectively. For this reason they are merged in AspX and GluX in [136].
3. Not all aminoacids are well recovered by the analysis, some are classified as “stable” and some as “unstable” (plus glycine because some glycine is produced in the decomposition of contaminating nucleic acids). In order to have a stable denominator, the aminoacid relative frequencies are expressed with respect to the sum of the stable aminoacids.
4. The two rare aminoacids Cys and Trp were not always detectable.

TO summarize, from [136] we have data for 14 individuals aminoacid and the AspX and GluX groups. There are 11 bacterial species plus *Tetrahymena pyriformis* which is not used in the regression analysis but added as an illustrative point. There are 22 rows in the dataset because there are 4 replicates for *Escherichia coli*, 3 for *Bacillus subtilis*, 2 for *B. cereus*, *Serratia marcesens*, *Sarcina lutea*, *Pseudomonas aeruginosa* and *Micrococcus lysodeikticus*. The following R code was used to re-create SUEOKA’s plots from [136].

```
NS61 <- read.table("local/NS61.csv", sep = "\t", header = TRUE, dec = ",")
sueoplot <- function(aa){
  n <- nrow(NS61)
  x <- NS61$GC[-n]
  iaa <- which(colnames(NS61) == aa)
```

¹Remember that at that time the genetic code wasn’t yet deciphered. As stated p. 1147 in the paper “[t]he present data seem to support universality of the code among bacteria. The presence of different codes among the bacteria would clearly preclude finding any correlation”. And since the results were consistent for *Tetrahymena pyriformis* that “[t]his may suggest that the underlying coding is also common to protozoa”. As far as I know, this is the first evidence of the fascinating universality of the genetic code. Moreover, SUEOKA’s results gave some clues for the structure of the genetic code since we expect aminoacid with a positive correlation to be encoded by GC-rich codons and those with a negative one by GC-poor codons.

²I was able to re-calculate the same slopes values except for methionine in figure 2.17 page 31 and for GluX in figure 2.35 page 63 for an unknown reason.

```

y <- NS61[, iaa]
ymax <- max(y, na.rm = TRUE)
y <- y[-n] # Remove Tetrahymena pyriformis
stbl <- ifelse(aa %in% colnames(NS61)[3:13], "(stable)", "(unstable)")
sunflowerplot(x, y, xlab = "Genomic GC content [%]", las = 1,
              ylab = "Aminoacid frequency [%]", main = paste(aa, stbl),
              xlim = c(20, 80), ylim = c(0, ymax), pch = 1)
abline(lm(y~x))
points(NS61[n, "GC"], NS61[n, iaa], pch = 16)
mtext(bquote(r^2 == .(signif(cor(x,y)^2, 2))), adj = 0)
mtext(bquote(alpha == .(signif(lm(y~x)$coef[2], 2))), adj = 1)
}

```

2.2.3 GC content as a nuisance parameter

A nuisance parameter is any parameter which is not of immediate interest but which *must* be accounted for in the analysis of those parameters which are of interest. To illustrate this, I will use a great example that was pointed to me by Thomas LUMLEY from the University of Washington on the R-help diffusion list³. The data [141, 117] interest is well described in [53]. It contains (among other variables) for 654 human individuals:

- **FEV**: a quantitative variable (in $l.s^{-1}$) which name is an acronym for “Forced Expiratory Volume”. This is the volume of air expelled after one second of constant effort, the higher the value is, the better for your health it is.
- **Smoker**: a binary qualitative variable stating whether the individual is a current smoker of cigarettes (**Current**) or not (**Non**).

THE following **R** code was used to download the data from the internet and to save it in a local file in XDR [139] format:

```

path <- "http://www.statsci.org/data/general/fev.txt"
fev <- read.table(path, sep = "\t", header = TRUE)
save(fev, file = "local/fev.Rda")

```

WE want to study the impact of smoking on respiratory performance. Let’s use the famous STUDENT’s *t*-test [40]:

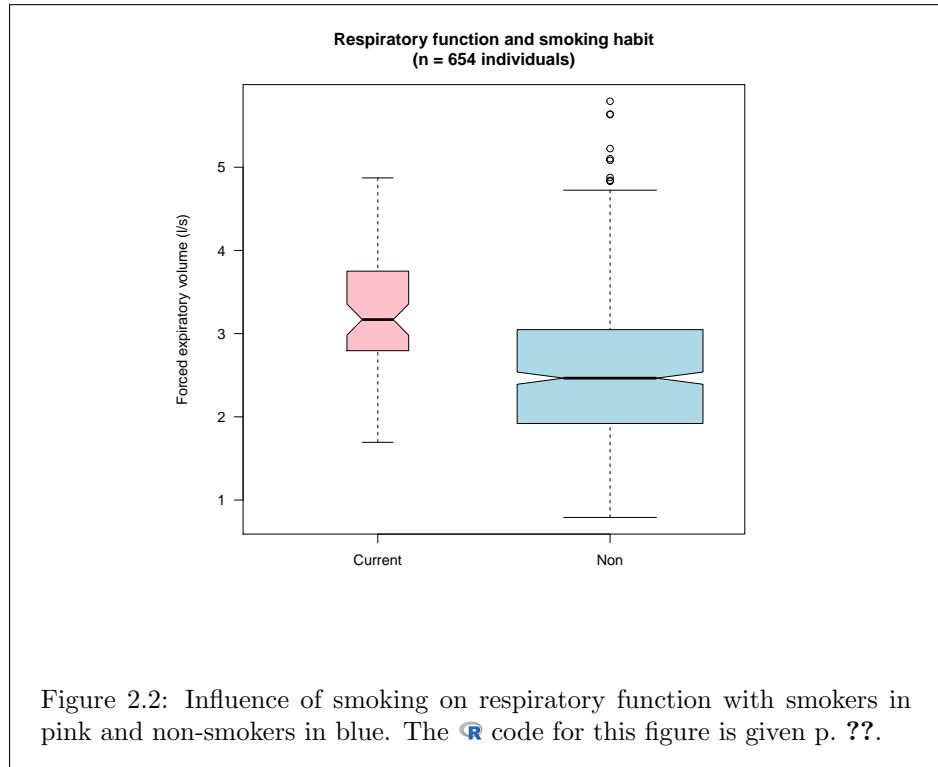
```

load("local/fev.Rda")
t.test(FEV~Smoker, data = fev)
      Welch Two Sample t-test
data:  FEV by Smoker
t = 7.1496, df = 83.273, p-value = 3.074e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5130126 0.9084253
sample estimates:
mean in group Current      mean in group Non
      3.276862              2.566143

```

AT any decent α critical level, we reject the null hypothesis, there is as expected an effect of smoking on respiratory performances. Figure 2.2 page 8 illustrate this in a different way: since the notches of the two boxplots do not overlap then, at a critical level of 5%, we can reject the null hypothesis stating that the medians of the respiratory function are the same between the two groups. But, wait a minute, the results are *better* in the smoking group! Smoking is of course not good for your health, so why the smoking group has better respiratory results? Exteremly wrong conclusions are at hand when a *nuisance parameter* is neglected, as it is the case here.

³I was unable to find an archiv of this thread.

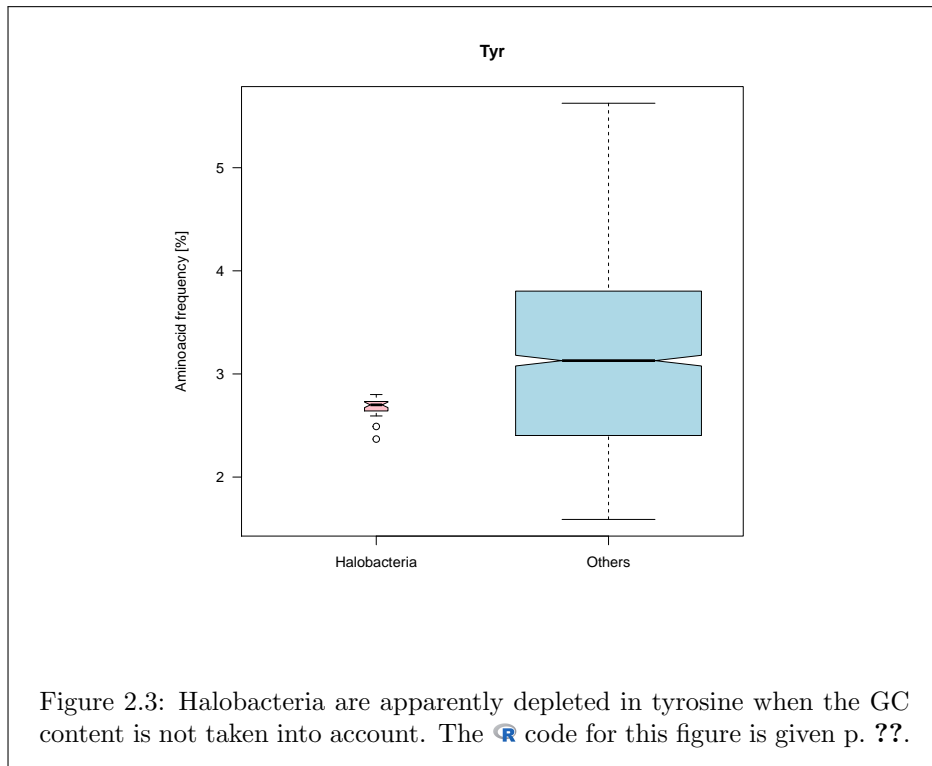


SPOILER alert! I don't want to ruin you the pleasure of discovering by yourself what is the *nuisance parameter* here. All the necessary information is present in the variables available in the dataset to figure out what's going on here. Just play with the data, so easy with R, before reading the following in a mirror. A simple summary (lev) will give you an hint by showing that there is a significant difference between the two groups. The age range is from 3 to 10 years, our individuals are in fact children. The age is a nuisance parameter because it has a strong effect on the forced expiratory volume as evidenced by with(lev, plot(Age, FEV)). All children that are less than 9 years old are non smokers, so that the non-smoking group is enriched by individuals with low FEV values, yielding to the spurious result when age is not taken into account.

LET'S give now a simple example of GC content as a nuisance parameter. Suppose that you are interested in the tyrosine content in Halobacteria because this class is known to have a special proteome content. Figure 2.3 page 9 shows that there is an extremely significant depletion of tyrosine in Halobacteria. This is completely wrong since they are in the opposite enriched in this aminoacid as we will see in section 2.3.4 page 25 where the GC content is taken into account. That's why SUEOKA's plot is so important.

2.2.4 Aminoacid frequencies under neutral conditions

THE function giving aminoacid frequencies as function of GC content, θ , in purely neutral conditions [70] is given for the standard genetic code by:



$$P(\theta, aa) = \frac{f(\theta)}{8 - (1 - \theta)^2(1 + \theta)} \quad (2.7)$$

where $f(\theta)$ is different from aminoacid to aminoacid:

$$f(\theta) = \begin{cases} (1 - \theta)^2(2 - \theta) & \text{if } aa \in \{\text{Ile}\} \\ (1 - \theta)^2 & \text{if } aa \in \{\text{Phe, Lys, Tyr, Asn}\} \\ 1 - \theta^2 & \text{if } aa \in \{\text{Leu}\} \\ (1 - \theta)^2\theta & \text{if } aa \in \{\text{Met}\} \\ (1 - \theta)\theta & \text{if } aa \in \{\text{Asp, Glu, His, Gln, Cys}\} \\ 2(1 - \theta)\theta & \text{if } aa \in \{\text{Val, Thr}\} \\ 3(1 - \theta)\theta & \text{if } aa \in \{\text{Ser}\} \\ (1 - \theta)\theta^2 & \text{if } aa \in \{\text{Trp}\} \\ \theta(\theta + 1) & \text{if } aa \in \{\text{Arg}\} \\ 2\theta^2 & \text{if } aa \in \{\text{Gly, Pro, Ala}\} \end{cases} \quad (2.8)$$

The corresponding `R` code is:

```

aatho <- function(the, aa){
  den <- 8 - (1 - the)^2*(1 + the)
  if(aa %in% c("Ile")) return(((2 - the)*(1 - the)^2)/den)
  if(aa %in% c("Phe", "Lys", "Tyr", "Asn")) return((1 - the)^2/den)
  if(aa %in% c("Leu")) return((1 - the^2)/den)
  if(aa %in% c("Met")) return((the*(1 - the)^2)/den)
  if(aa %in% c("Asp", "Glu", "His", "Gln", "Cys")) return((the*(1 - the))/den)
  if(aa %in% c("Val", "Thr")) return((2*the*(1 - the))/den)
  if(aa %in% c("Ser")) return((3*the*(1 - the))/den)
}

```

```

if(aa %in% c("Trp")) return(((1 - the)*the^2)/den)
if(aa %in% c("Arg")) return(((1 + the)*the)/den)
if(aa %in% c("Gly", "Pro", "Ala")) return((2*the^2)/den)
stop("unknown aa")
}

```

This model defines three classes of aminoacids:

```

classaa <- function(aa){
  if(nchar(aa) == 1) aa <- aaa(aa)
  # 6 aa decreasing with GC content:
  if(aa %in% c("Ile", "Phe", "Lys", "Tyr", "Asn", "Leu")) return(1)
  # 4 aa increasing with GC content:
  if(aa %in% c("Gly", "Pro", "Ala", "Arg")) return(3)
  # 10 aa poorly affected by GC content:
  return(2)
}

```

FIGURE 2.4 page 11 shows the distribution of aminoacid is far from uniform and poorly explained by the number of codons per aminoacid. There are clearly selective constraints here. Here is an utility function to explore aminoacid frequencies:

```

showaa <- function(aalist){
  x <- tdd$tdgc
  y <- rowSums(cbind(tdd[, which(colnames(tdd) %in% aalist)], 0))
  plot(x, y, xlim = c(0, 100), ylim = c(0, max(y)), las = 1,
       xlab = "GC content [%]", ylab = "Aminoacid content [%]",
       pch = 19, cex = tdd$cex, main = paste(aalist, collapse = " "), col = col2alpha("black", 0.25))
  isa <- which(tdd$superkingdom == 2157)
  points(x[isa], y[isa], pch = 21, bg = col2alpha("red", 0.8), cex = tdd$cex[isa])
  ish <- which(tdd$class == 183963) # Halobacteria
  points(x[ish], y[ish], pch = 21, bg = col2alpha("orange", 0.8), cex = tdd$cex[ish])
  ishq <- which(tdd$genus == 293431)
  points(x[ishq], y[ishq], pch = 21, bg = col2alpha("yellow", 0.8), cex = tdd$cex[ishq])
  abline(lm(y~x), lty = 2)
  xx <- seq(0, 100, le = 256)
  aath <- rep(0, 256)
  for(aa in aalist) aath <- aath + sapply(xx/100, function(x) 100*aatheo(x, aa))
  lines(xx, aath, type = "l")
  legend("bottomleft", inset = 0.02, legend = c("Archaea - not Halobacteria",
        "Halobacteria - not Haloquadratum spp.",
        "Haloquadratum spp.", "Eubacteria"), pch = 21,
        pt.bg = c(col2alpha("red", 0.8), col2alpha("orange", 0.8),
        col2alpha("yellow", 0.8), col2alpha("black", 0.25)),
        bg = grey(0.9))
  mtext(bquote(r^2 == .(signif(cor(x,y)^2, 3))), adj = 0)
  mtext(bquote(alpha == .(signif(lm(y~x)$coef[2], 3))), adj = 1)
  legend <- c("Linear fit", "Neutral model")
  legend("bottomright", inset = 0.02, bg = grey(0.9), lty = c(2, 1), legend = legend)
}

```

This code is used to generate all figures:

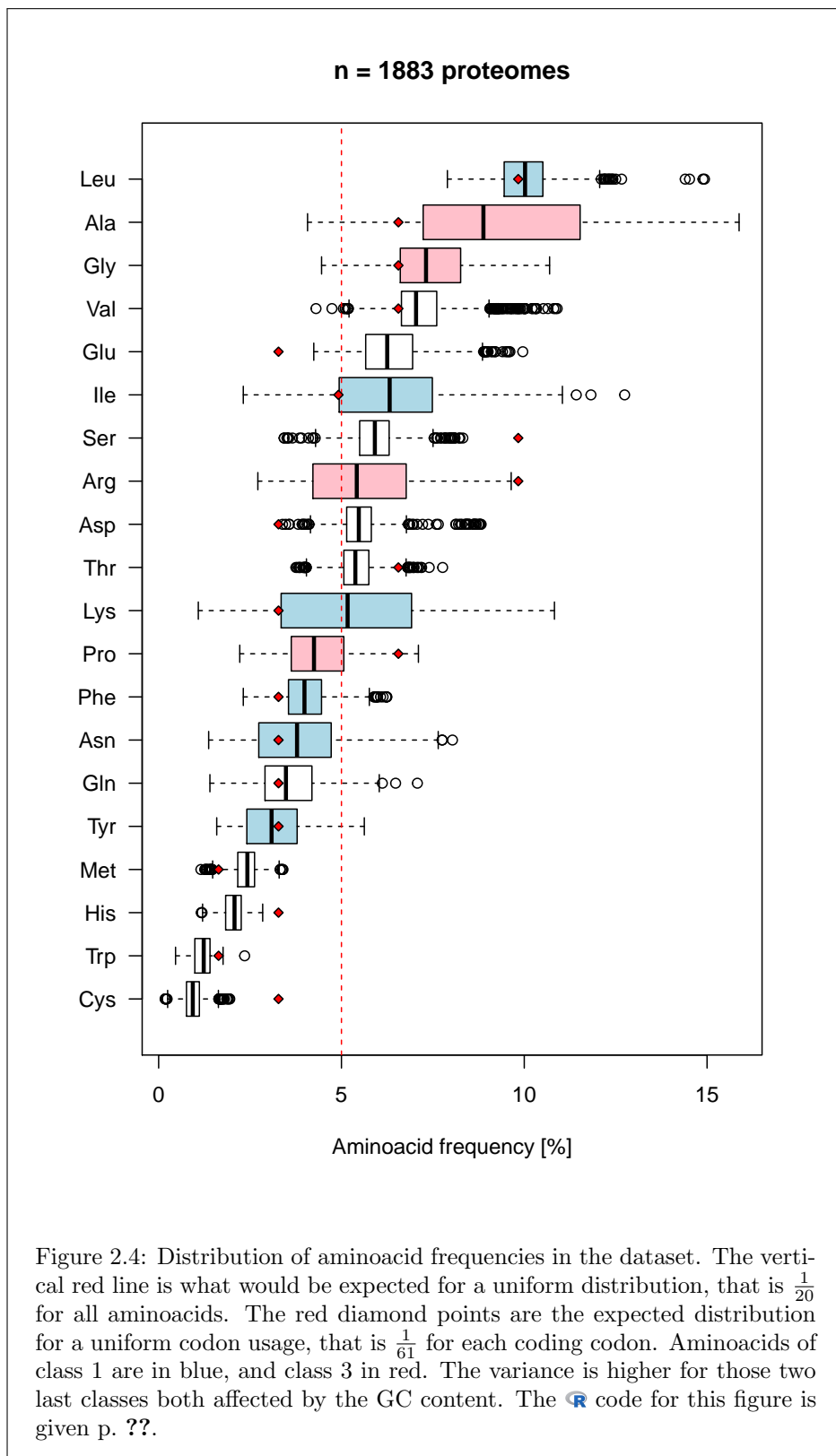
```

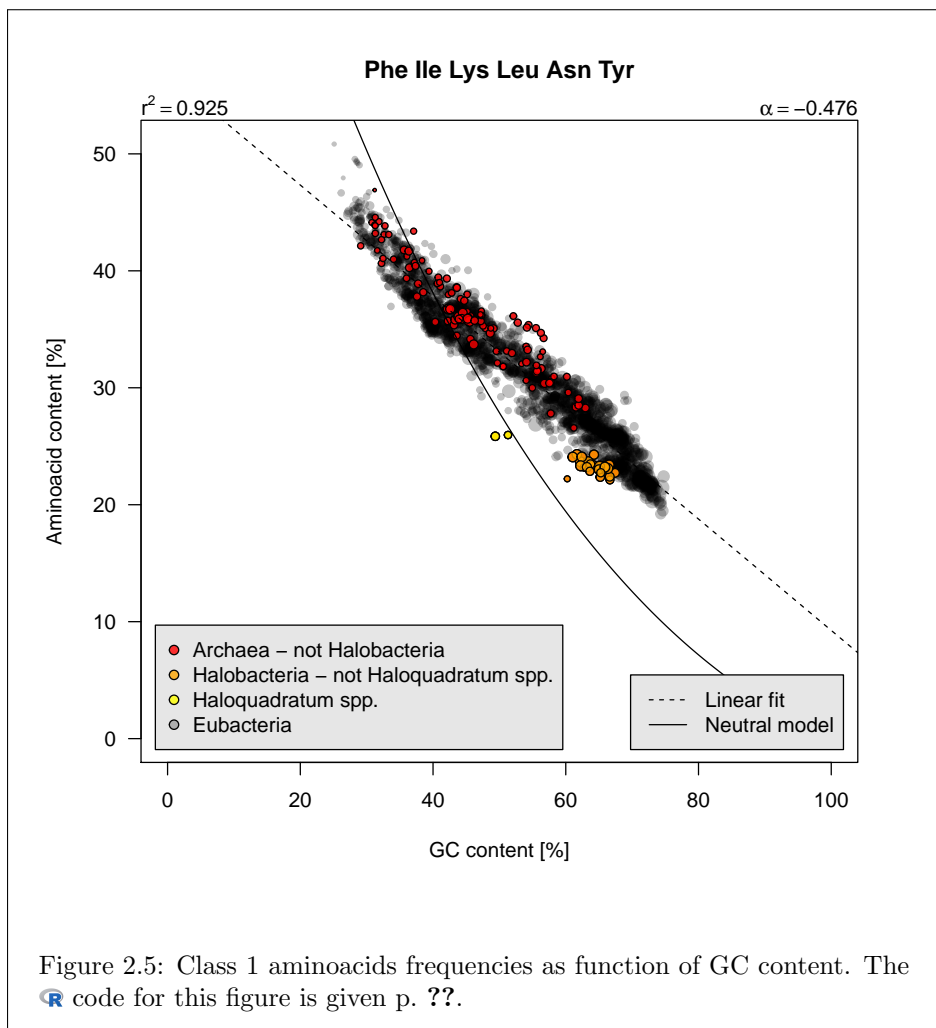
todo <- aaa()[-1]
for(i in todo){
  fname <- paste("figs/auto-", i, ".pdf", sep = "")
  pdf(fname)
  showaa(i)
  dev.off()
}

```

2.2.5 Aminoacid classes with respect to GC content

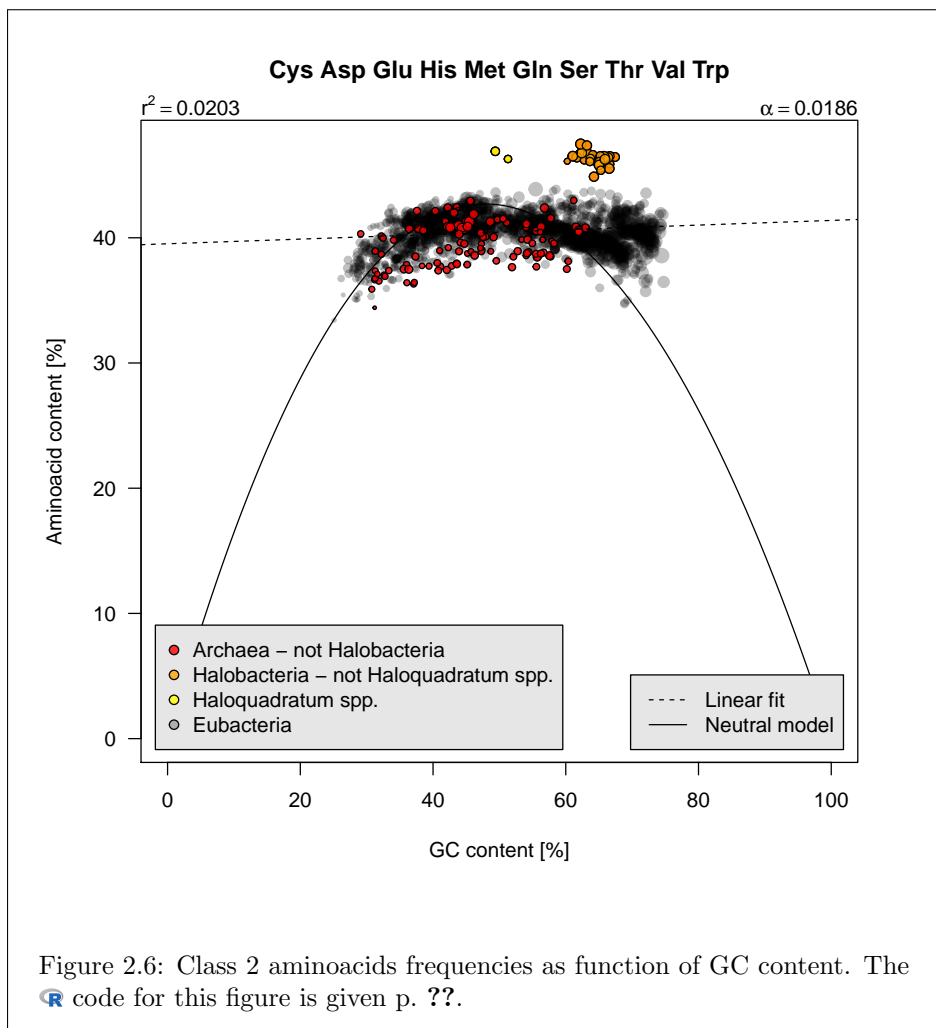
THE model 2.8 defines three classes of aminoacids with respect to their expected behaviour to genomic GC content:

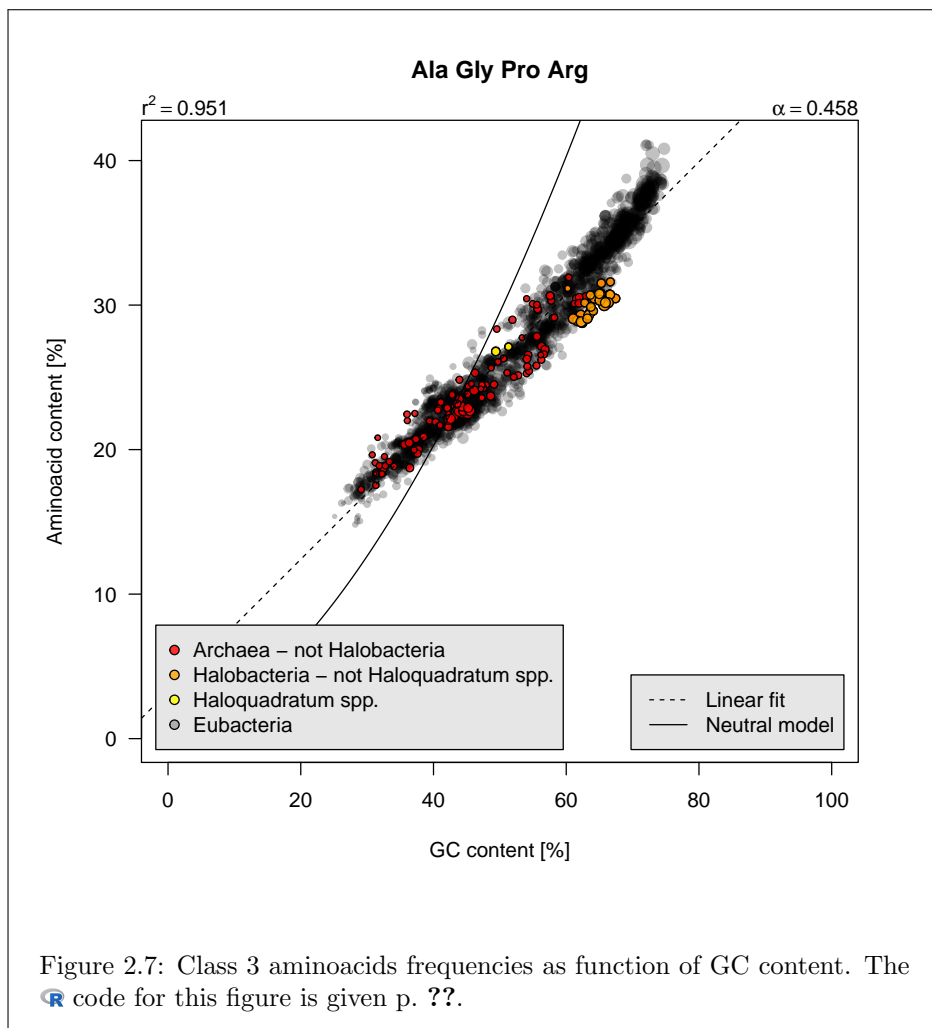


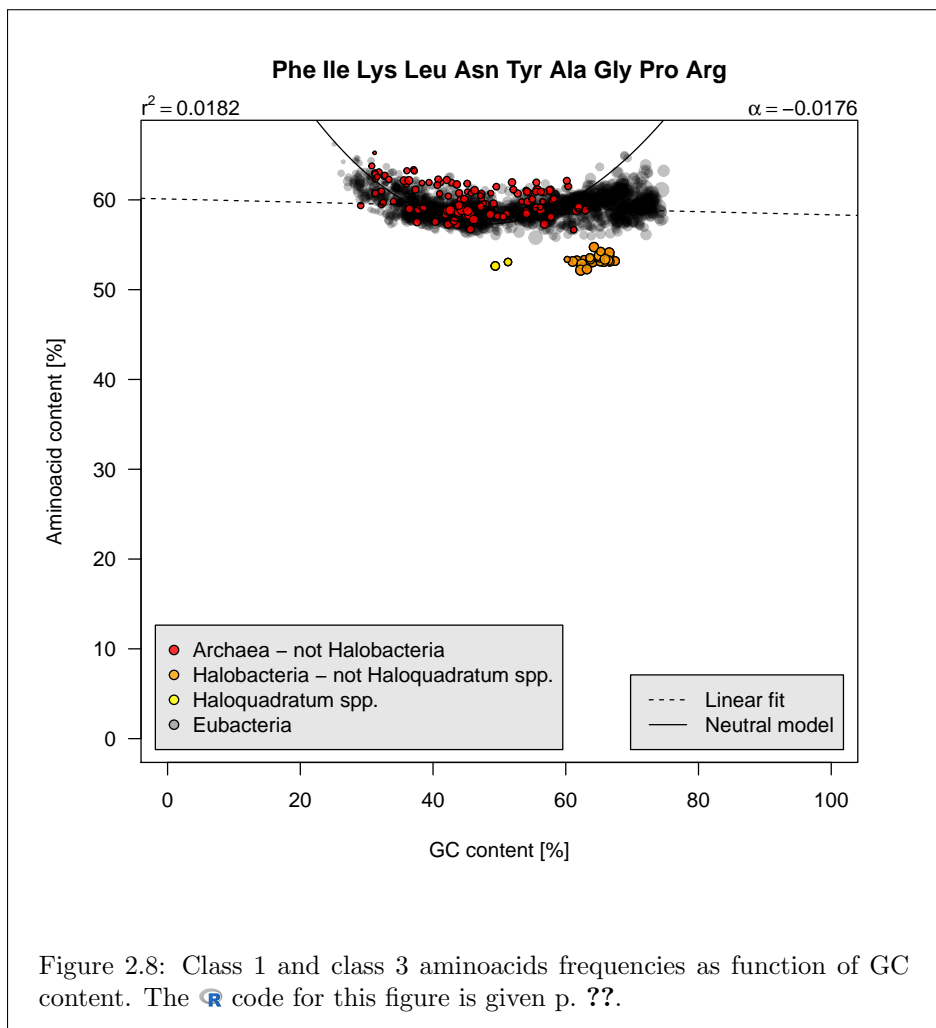


1. Six aminoacids whose frequencies are expected to decrease with GC content. Figure 2.5 page 12 shows that they represent about 45% of aminoacids in low-GC bacteria and about 20% in high-GC bacteria. On average, an increase of 10% for the GC content will decrease their frequency by 5%.
2. Ten aminoacids whose frequencies are poorly affected by GC content. Figure 2.6 page 14 shows that they represent about 40% of aminoacids, that is slightly less than the 50% we would expect from a uniform distribution.
3. Four aminoacids whose frequencies are expected to increase with GC content. Figure 2.7 page 15 shows that they represent about 15% of aminoacids in low-GC bacteria and about 40% in high-GC bacteria. On average, an increase of 10% for the GC content will increase their frequency by 5%.

FIGURE 2.8 page 16 shows that the decrease in class 1 is compensated by the increase in class 3 (this is logical since class 2 is almost constant and the grand total 100% by construction) so that these 10 aminoacids represent about 60% of aminoacids, that is slightly more than the 50% we would expect from a uniform distribution.







2.3 Class 1 aminoacids

2.3.1 Isoleucine

ISOLEUCINE is a non-polar uncharged aliphatic aminoacid encoded by three codons. It is highly sensitive to the GC content since its frequency decreases from 12% in low-GC bacteria to 2% in high-GC bacteria, that is a factor 6. Figure 2.9 page 18 shows that the results are consistent with [136]. The linear model summarises well the general trend ($r^2 \approx 0.9$) but the distribution of residual is non-random, with a sigmoidal shape. There is a small trend for low-GC archaea to use more Ile than eubacteria. There is also a small trend for halobacteria to use less Ile than eubacteria. Its frequency is close to what would be expected from uniform codon usage at GC below 50% but higher at GC above 50% as if there were a selective pressure to maintain a minimal frequency. The two top outliers are:

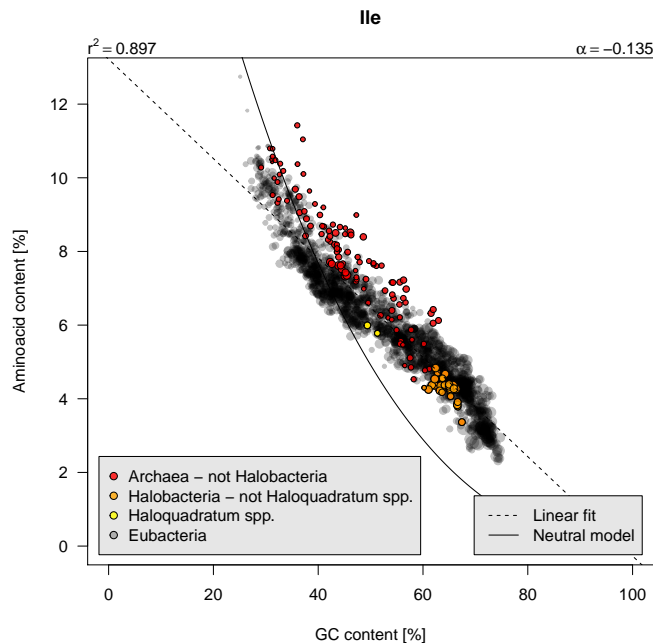
```
tdd[tdd$Ile > 11.5, "organism"]
[1] "buchnera_aphidicola"      "wigglesworthia_glossinidia"
```

THE two outliers, *Buchnera aphidicola* [87] and *Wigglesworthia glossinidia* [1], are both low-GC small genomes endosymbiont bacteria. Their Ile frequency is above 11.5%. The two following outliers are:

```
tdd[tdd$Ile < 11.5 & tdd$Ile > 11, "organism"]
[1] "ignisphaera_aggregans" "picrophilus_torridus"
```

THE two following outliers, *Picrophilus torridus* [124] and *Ignisphaera aggregans* [91], are both acidophilic (pH 0.7 and 6.4 respectively) archaeae with high T_{opt} (60°C and 92°C, respectively).

Could this be a consequence of the deregulation of aminoacid biosynthetic pathways?



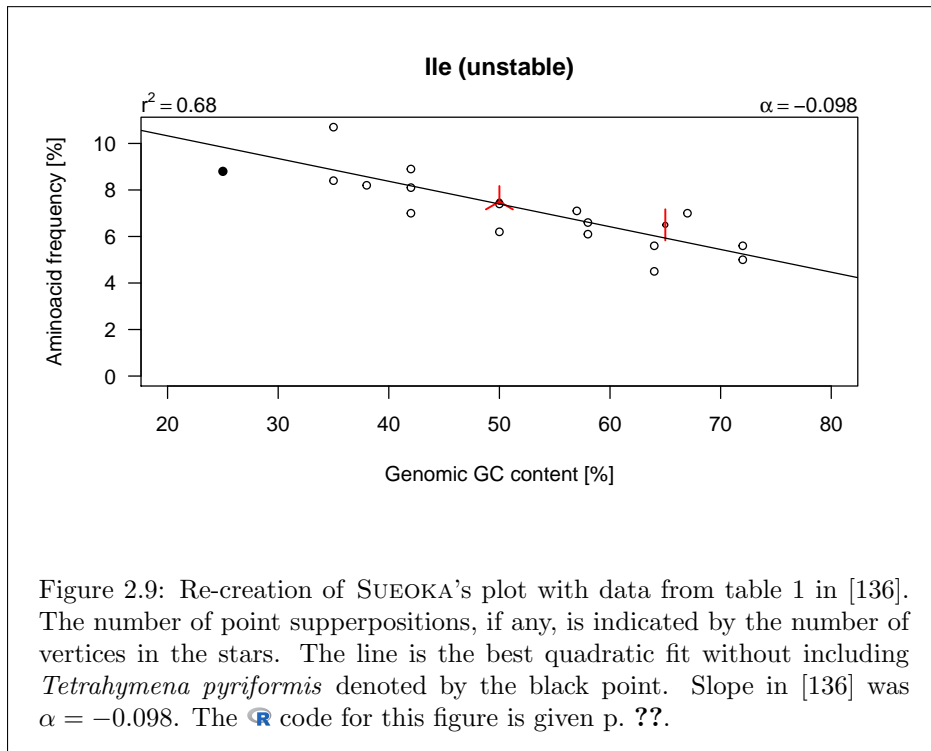


Figure 2.9: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = -0.098$. The `R` code for this figure is given p. ??.

2.3.2 Phenylalanine

PHENYLALANINE is an aromatic neutral nonpolar aminoacid encoded by two codons. It is sensitive to the GC content since its frequency decreases from 6% in low-GC bacteria to 2% in high-GC bacteria, that is a factor 3. Figure 2.10 page 20 shows that the results are consistent with [136]. The linear model summarises well the general trend ($r^2 \approx 0.7$) but the distribution of residual is non-random, with a sigmoidal shape. Its frequency is close to what would be expected from uniform codon usage at GC below 40% but higher at GC above 40% as if there were a selective pressure to maintain a minimal frequency. Most low-Phe are archaeae but not all archaeae are low-Phe. The top-outlier are from two genera (*Borrelia* = *Borrelia* and *Campylobacter*):

```
tdd[tdd$Phe > 5.8, "organism"]
```

```
[1] "borrelia_afzelii"           "borrelia_burgdorferi"
[3] "borrelia_garinii"         "borrelia_bavariensis"
[5] "borrelia_spielmanii"      "campylobacter_coli"
[7] "campylobacter_hominis"    "campylobacter_insulaenigrae"
[9] "campylobacter_jejuni"     "campylobacter_lari"
[11] "campylobacter_upsaliensis" "campylobacter_ureolyticus"
[13] "campylobacter_volucris"
```

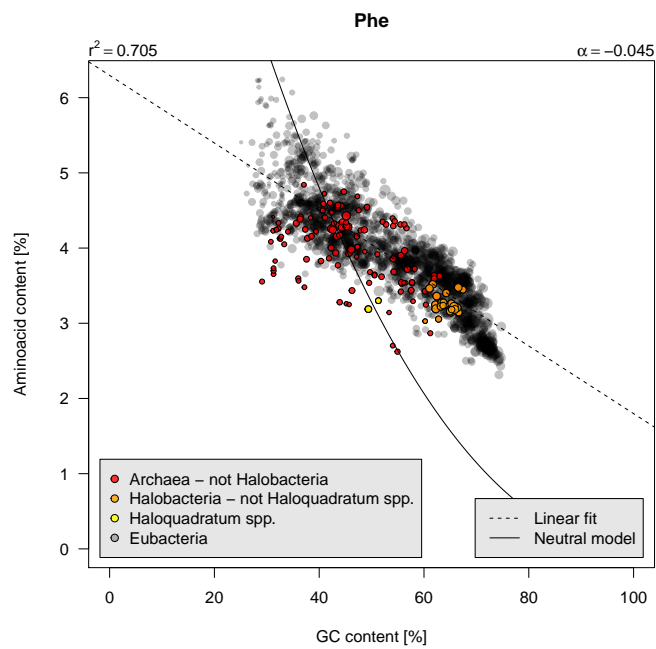
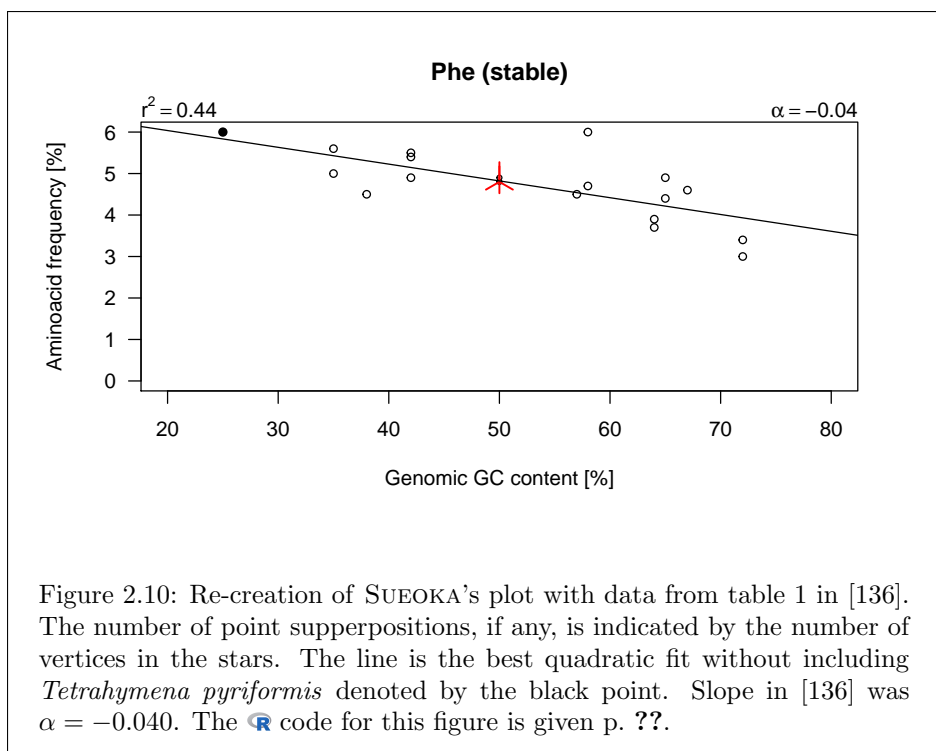
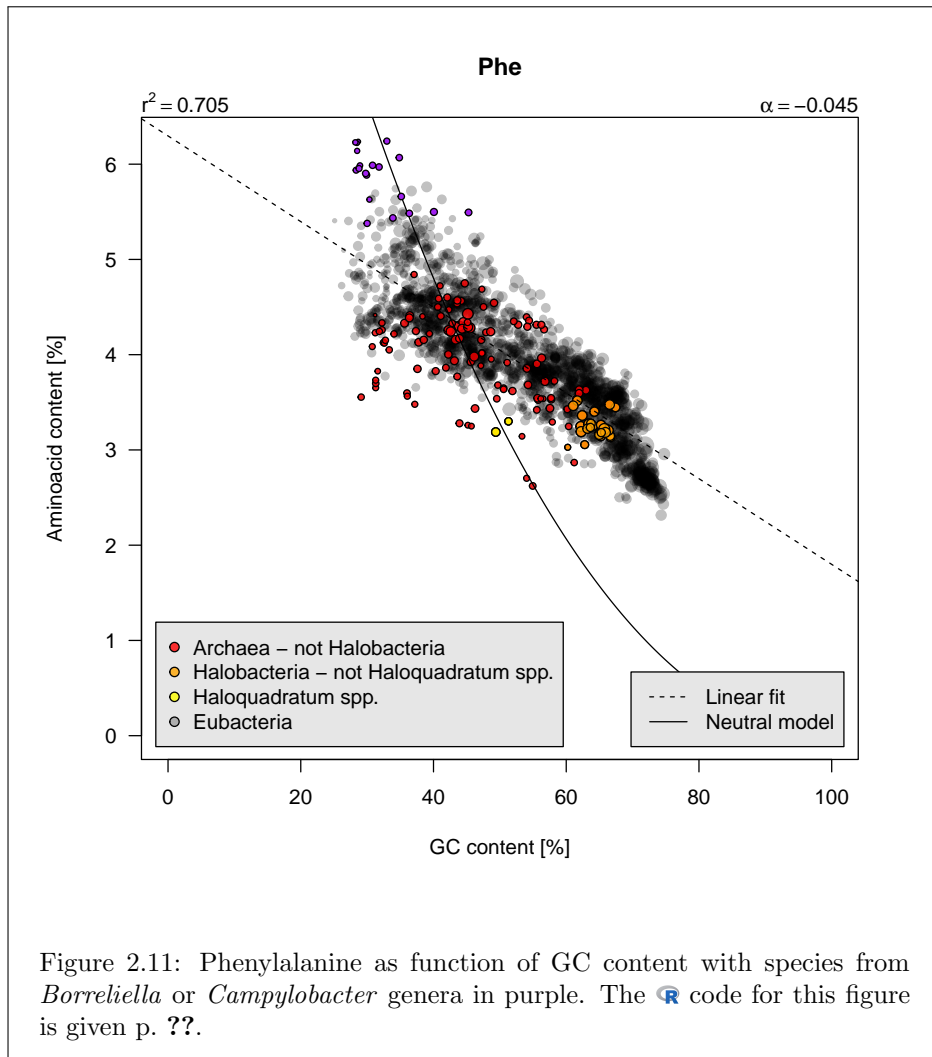


FIGURE 2.11 page 21 shows that the species from these two genera are generally rich in Phe, however the observed frequencies are not anomalously high as compared to what is expected from a uniform codon usage. This is somewhat redundant to say that they have a GC-poor genome.



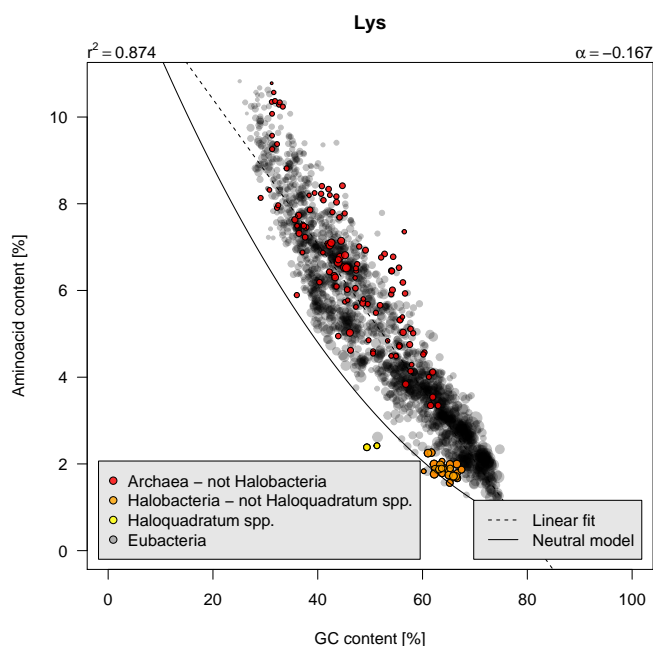


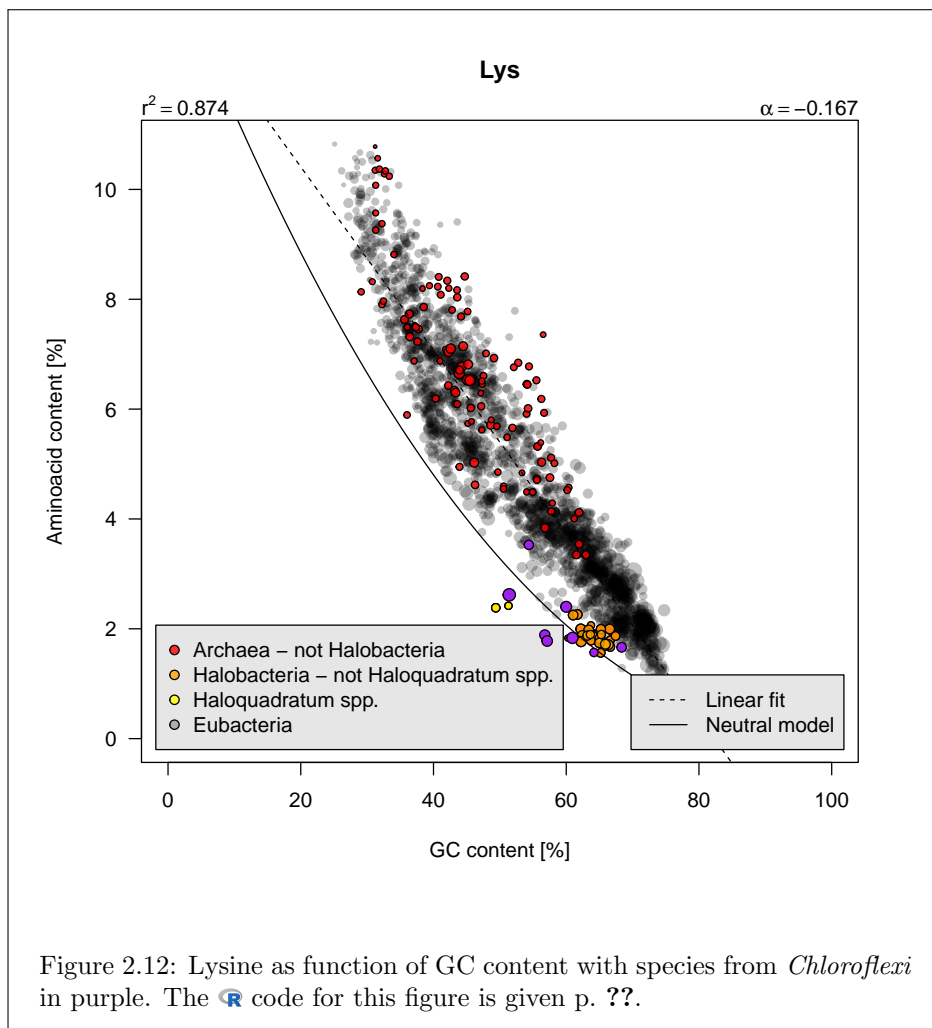
2.3.3 Lysine

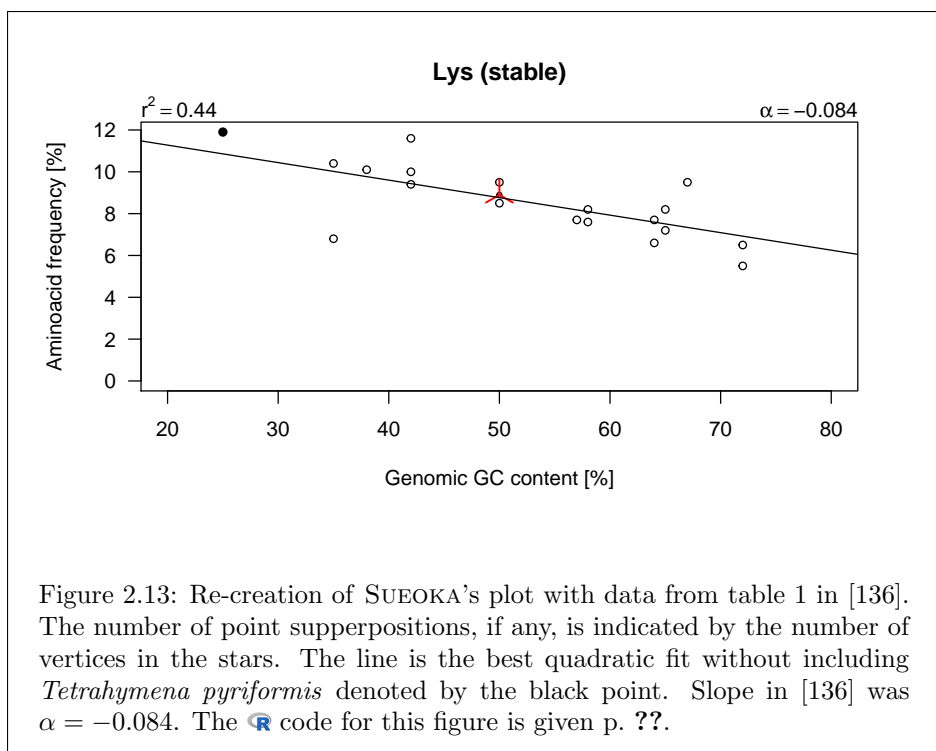
LYSINE is a basic charged aliphatic amino acid encoded by two codons. It is highly sensitive to the GC content since its frequency decreases from 10% in low-GC bacteria to 1% in high-GC bacteria, that is a factor 10. Figure 2.13 page 24 shows that the results are consistent with [136]. The linear model summarises well the general trend ($r^2 \approx 0.9$). Its frequency is close to what would be expected from uniform codon usage, but always higher as if there were a selective pressure to increase its frequency. Halobacteria tend to avoid this aminoacid, a known phenomenon [56]. The species with less Lys than predicted by the neutral model are:

```
tdd[100*sapply(tdd$tdgc/100, aatheo, aa = "Lys") > tdd$Lys, "organism"]
[1] "chloroflexus_aggregans"      "chloroflexus_aurantiacus"
[3] "haloquadratum_sp"           "haloquadratum_walsbyi"
[5] "halorubrum_sp"              "herpetosiphon_aurantiacus"
[7] "natrialba_magadii"          "roseiflexus_castenholzii"
[9] "thermomicrobium_roseum"
```

FROM those species, halobacteria are *Haloquadratum walsbyi*, *Halorubrum* sp. and *Natrialba magadii*. Remaining species are all *Chloroflexi* (Chlorobacteria) and figure 2.12 page 23 shows that there is a trend for species from this group to be depleted in Lys.

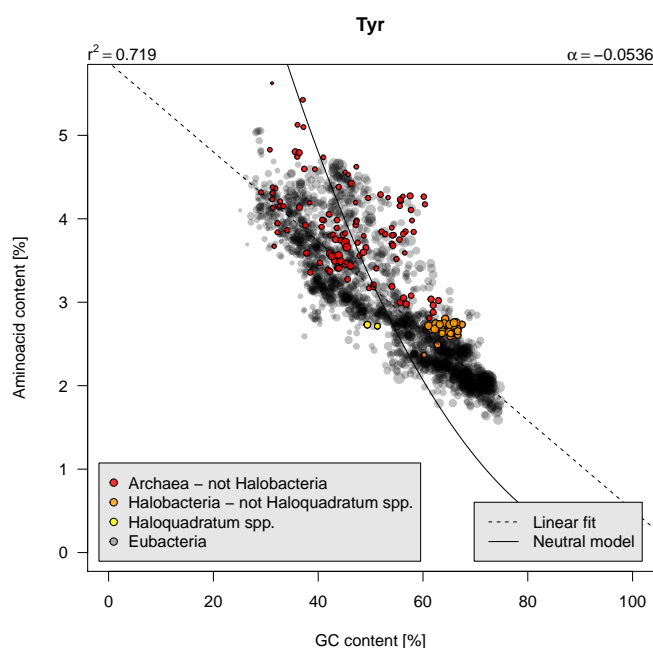






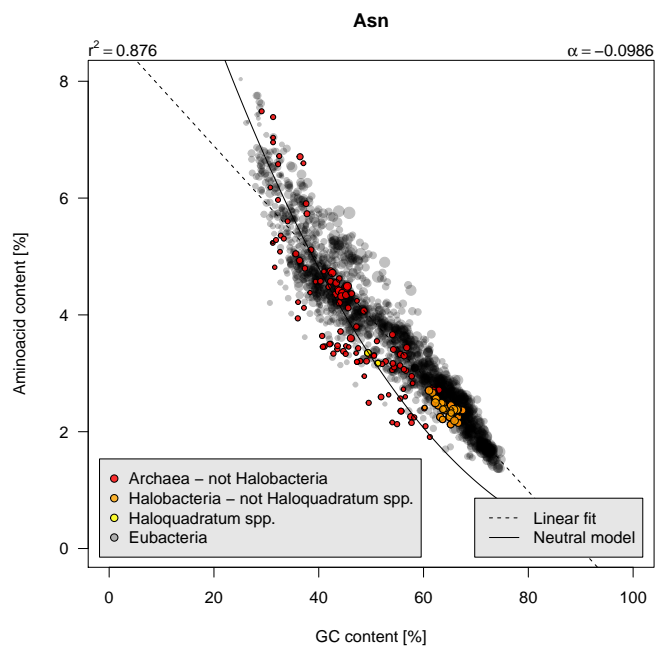
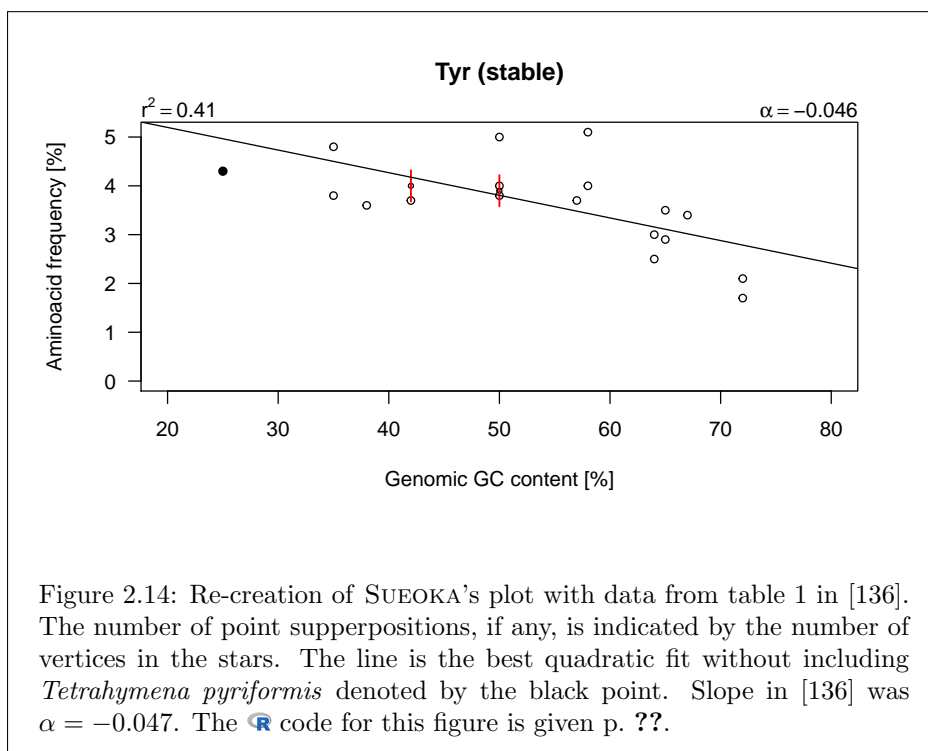
2.3.4 Tyrosine

TYROSINE is an aromatic polar hydrophobic aminoacid encoded by two codons. It is sensitive to the GC content since its frequency decreases from 5.5% in low-GC bacteria to 1.5% in high-GC bacteria, that is a factor 3.5. Figure 2.14 page 26 shows that the results are consistent with [136]. The linear model summarises well the general trend ($r^2 \approx 0.7$). Its frequency is close to what would be expected from uniform codon usage, but lower at GC below 50% and higher at GC above 50% as if there were a selective pressure to avoid too extreme values. Halobacteria tend to favor Tyr but there are two exceptions: *Haloquadratum walsbyi* [21] and *Halorubrum* sp.



2.3.5 Asparagine

ASPARAGINE is a polar aliphatic aminoacid encoded by two codons. It is very sensitive to the GC content since its frequency decreases from 8% in low-GC bacteria to 1% in high-GC bacteria, that is a factor 8. The linear model summarises well the general trend ($r^2 \approx 0.9$). Its frequency is close to what would be expected from uniform codon usage, but higher at GC above 50% as if there were a selective pressure to avoid a too low value. Halobacteria tend to avoid lightly Asn.

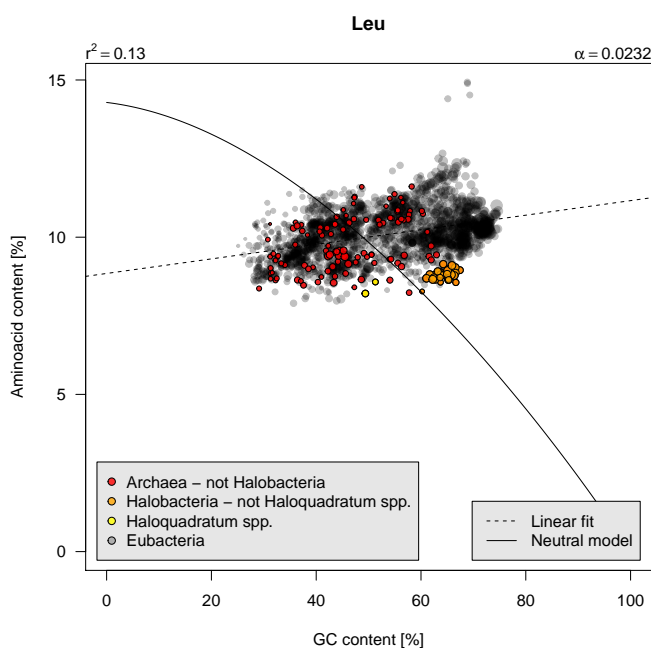


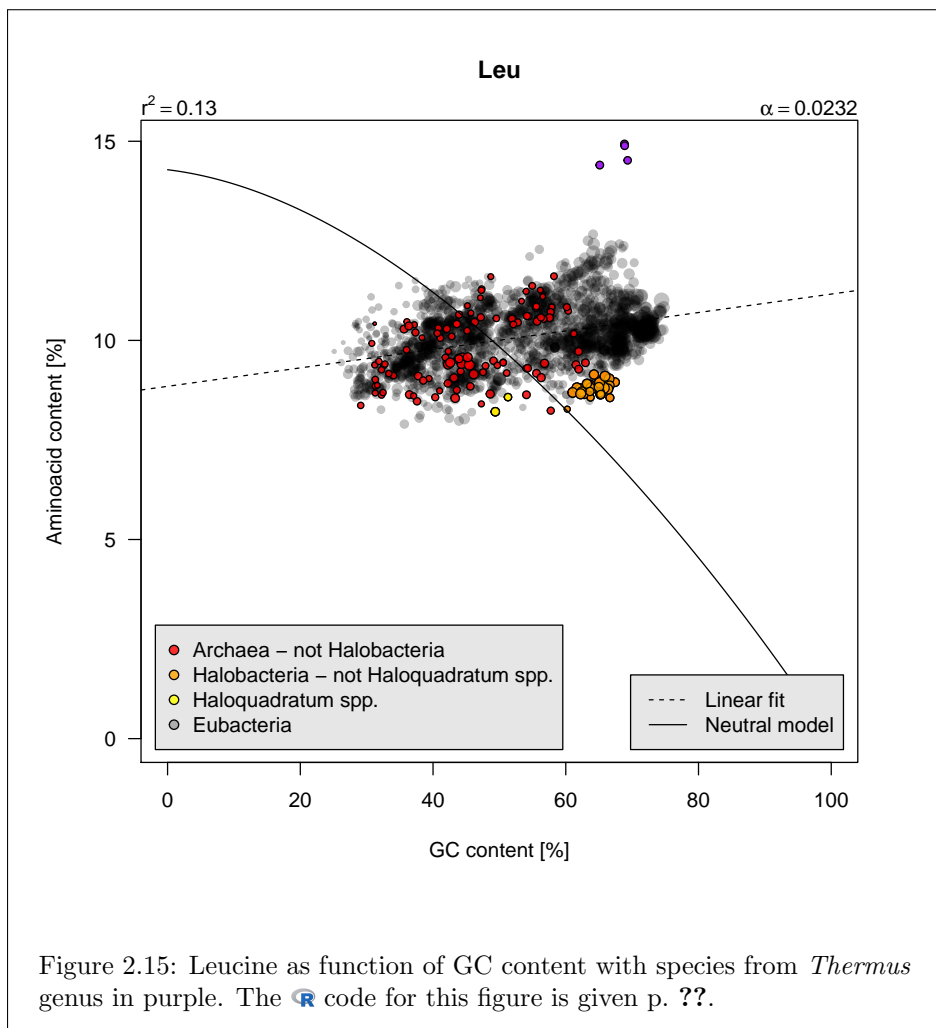
2.3.6 Leucine

LEUCINE is a non-polar aliphatic aminoacid encoded by 6 codons. It is supposed to decrease with GC but increase a little from 9.5% in low-GC bacteria to 10.5% in high-GC bacteria, that is a factor 1.1. Figure 2.16 page 29 shows that the results are consistent with [136]. The linear model summarises poorly the general trend ($r^2 \approx 0.14$). Halobacteria tend to avoid this aminoacid. The top-outliers are:

```
tdd[which(tdd[, "Leu"] > 14), c("organism", "domain", "topt", "Leu")]
  organism domain topt Leu
2175 thermus_oshimai Bacteria 65 14.92963
2176 thermus_scotoductus Bacteria 66 14.40196
2177 thermus_sp Bacteria 67 14.88733
2178 thermus_thermophilus Bacteria 75 14.52119
```

THE four outliers are all thermophilic eubacteria from the *Thermus* genus (*viz.* *T. oshimai* [154] *T. scotoductus* [63], *T. thermophilus* [97, 155], *T. sp.* [17]) and figure 2.15 page 28 shows that all species from this genus have a high Leu content. This is interesting because if this high-Leu frequency is an adaptation for high temperature, then there is no universal response for thermophily on aminoacid content.





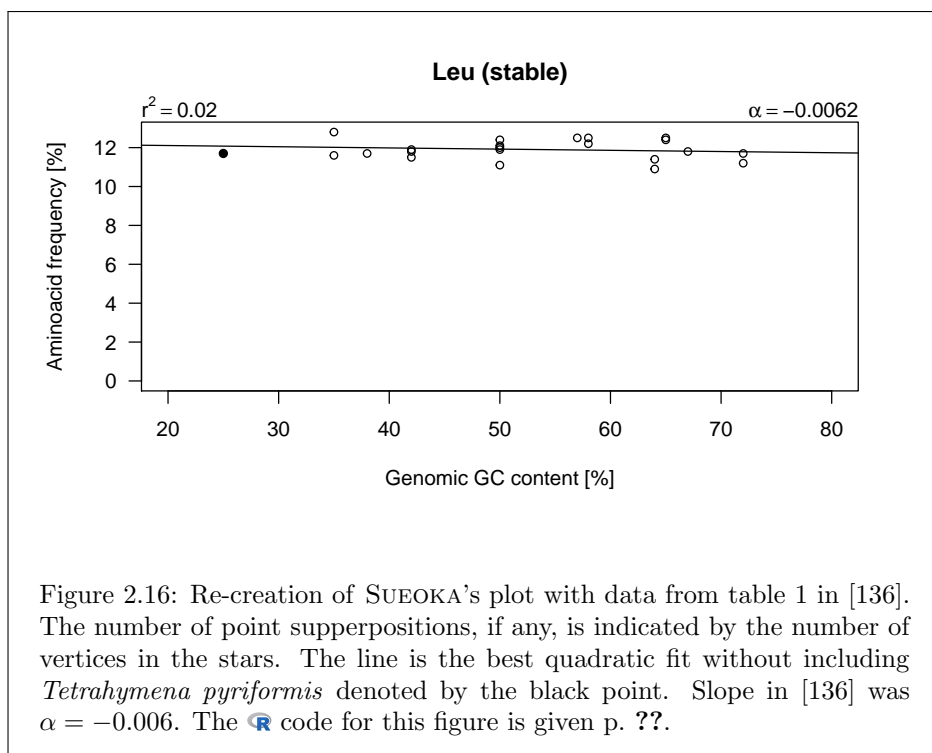

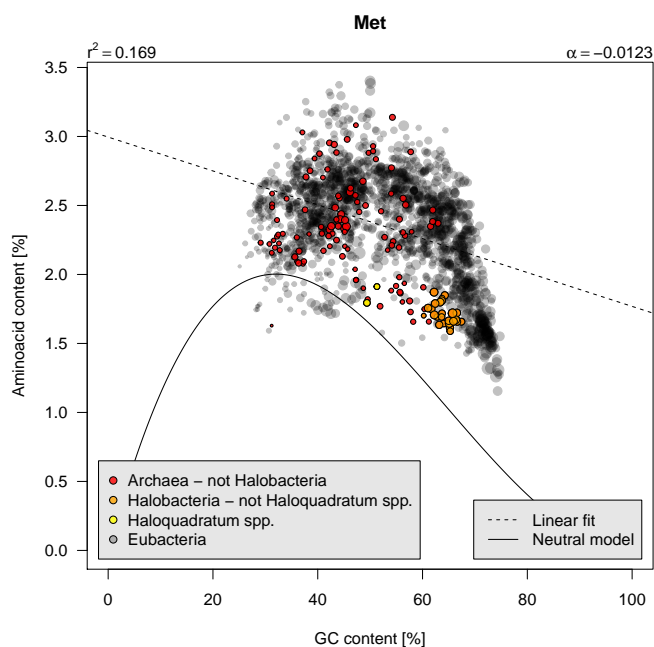


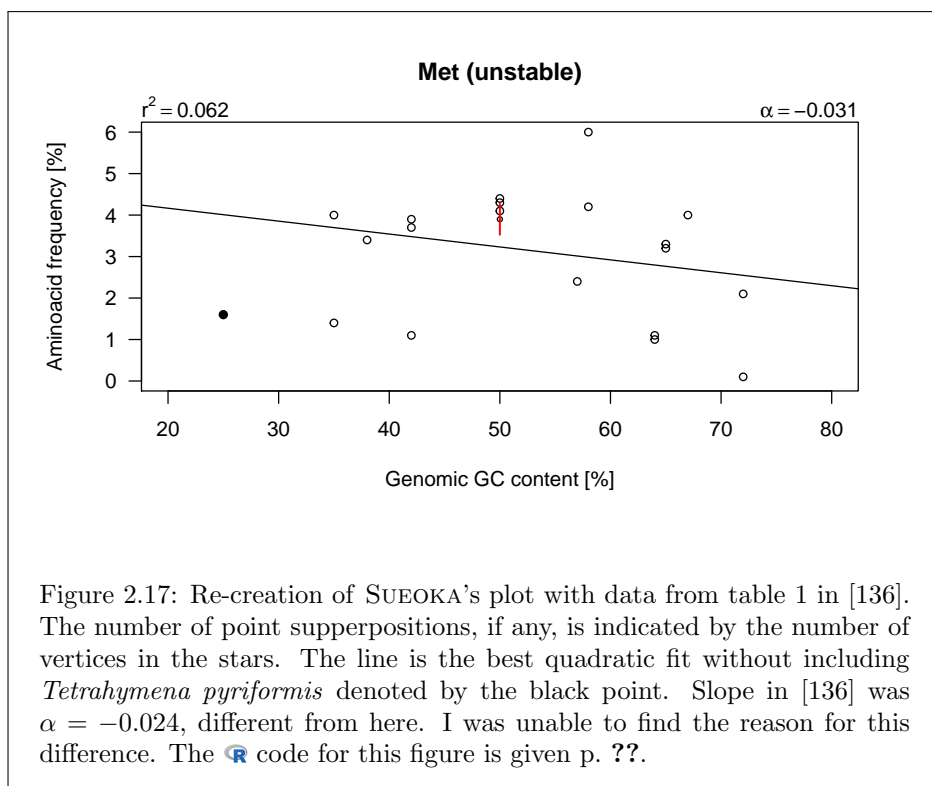
Figure 2.16: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = -0.006$. The  code for this figure is given p. ??.

2.4 Class 2 aminoacids

2.4.1 Methionine

METHIONINE is nonpolar aliphatic aminoacid encoded by a single codon. It is poorly sensitive to the GC content with frequency ranging from 1.5% to 3.5%, always above about 1% of what would be expected from uniform codon usage. It's only for GC greater than 50% that a decrease is observed. The linear model summarises poorly the general trend ($r^2 \approx 0.2$). Figure 2.17 page 31 shows that the results are consistent with [136]. Halobacteria tend to avoid Met.



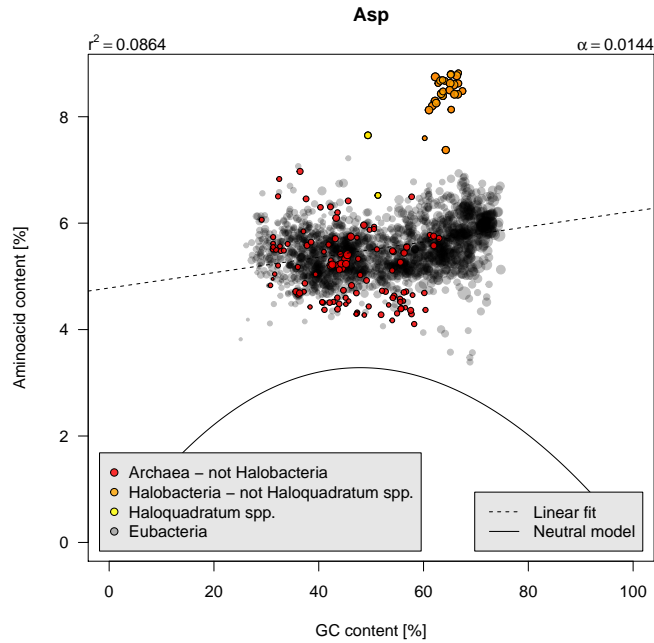


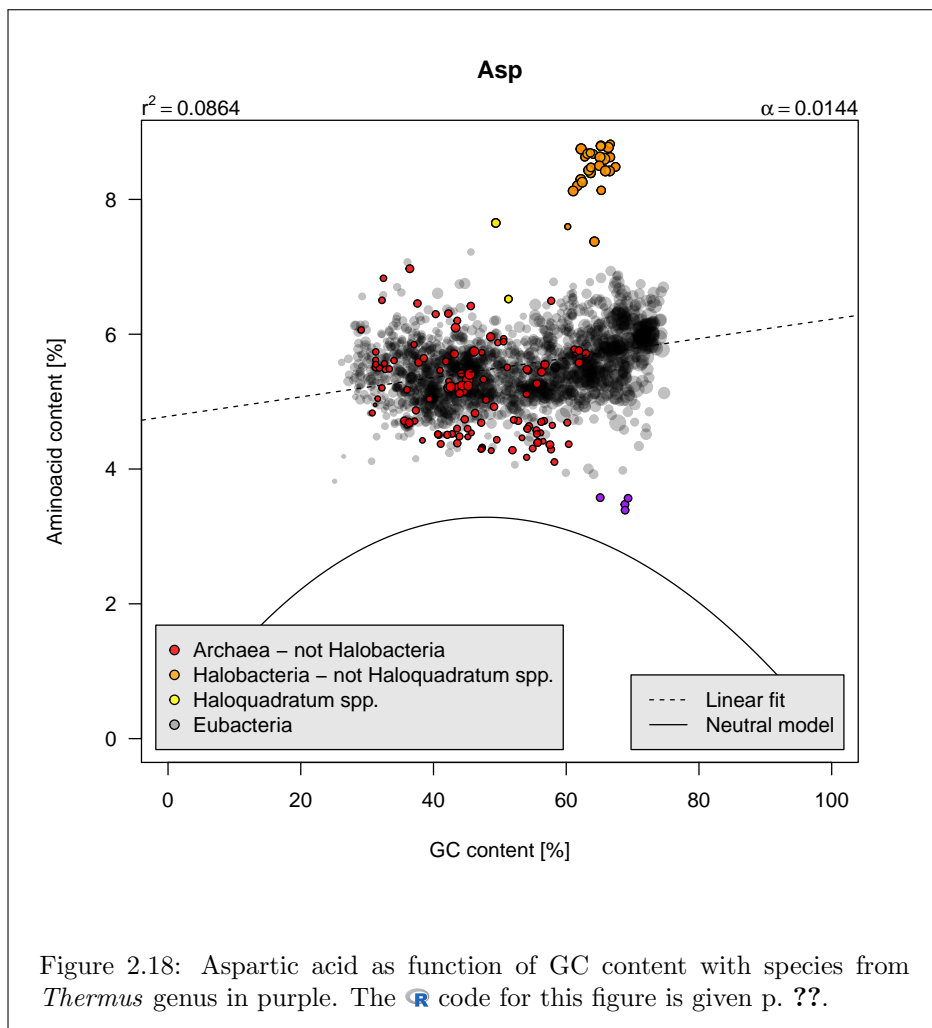
2.4.2 Aspartic acid

ASPARTIC acid is an acidic aminoacid encoded by two codons. It is almost unaffected by the GC content with an average concentration of 5.1% in low-GC bacteria to 5.9% in high-GC bacteria, that is a factor 1.15. The linear model summarises poorly the general trend ($r^2 \approx 0.1$). Its frequency is above about 2% of what would be expected from uniform codon usage. Halobacteria are highly enriched in Asp with a frequency close to 8%, a known phenomena [56]. The bottom outliers are:

```
tdd[tdd$Asp < 3.7, "organism"]
[1] "thermus_oshimai"      "thermus_scotoductus"  "thermus_sp"
[4] "thermus_thermophilus"
```

THE four outliers are all thermophilic eubacteria from the *Thermus* genus (*viz.* *T. oshimai* [154] *T. scotoductus* [63], *T. thermophilus* [97, 155], *T. sp.* [17]) and figure 2.18 page 33 shows that all species from this genus have a low Asp content. This is interesting because if this low-Asp frequency is an adaptation for high temperature, then there is no universal response for thermophily on aminoacid content.





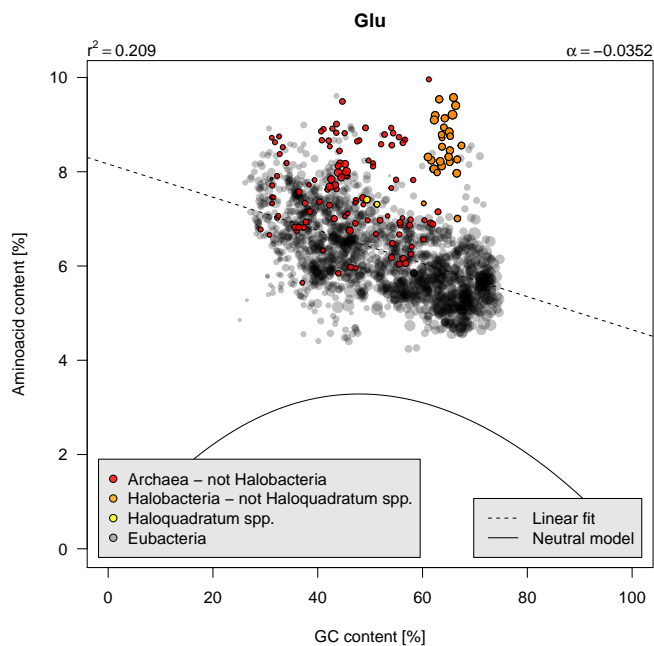
2.4.3 Glutamic acid

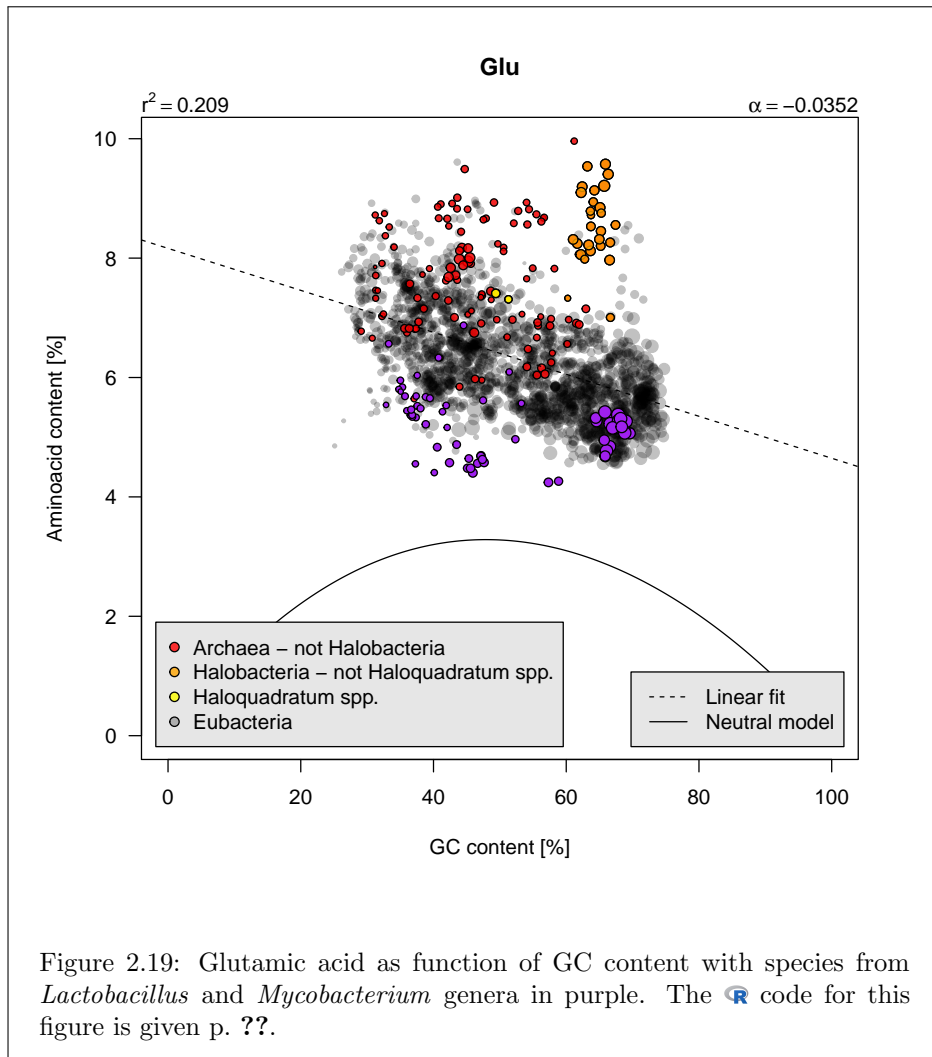
GLUTAMIC acid is an acidic aminoacid encoded by two codons. It is almost unaffected by the GC content with an average concentration of 7.2% in low-GC bacteria to 5.5% in high-GC bacteria, that is a factor 1.3. The linear model summarises poorly the general trend ($r^2 \approx 0.2$). Its frequency is above about 3-5% of what would be expected from uniform codon usage. Halobacteria are highly enriched in Glu with a frequency close to 9%, a known phenomena [56]. The bottom outliers are:

```
tdd[tdd$Glu < 4.3, "organism"]
```

```
[1] "lactobacillus_shenzhenensis" "mycobacterium_leprae"
```

THE two outliers are *Lactobacillus shenzhenensis* and *Mycobacterium leprae* but figure 2.19 page 35 shows that this is not a general trend for species from these genera.



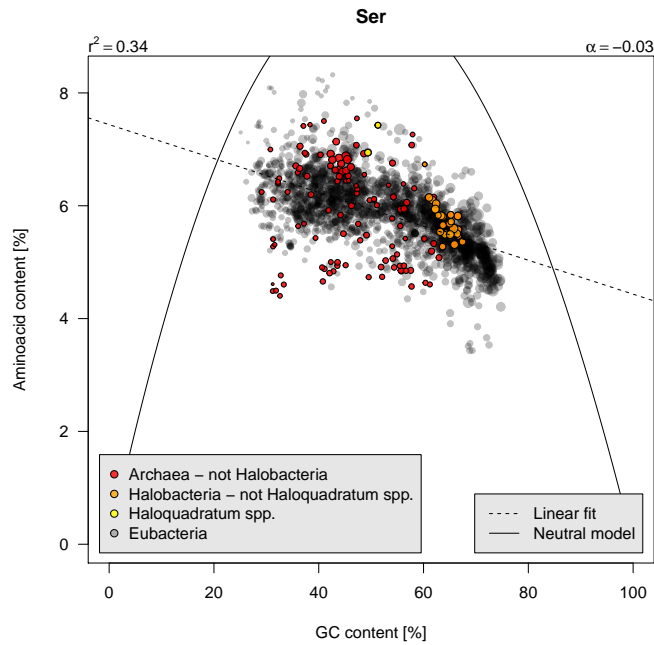


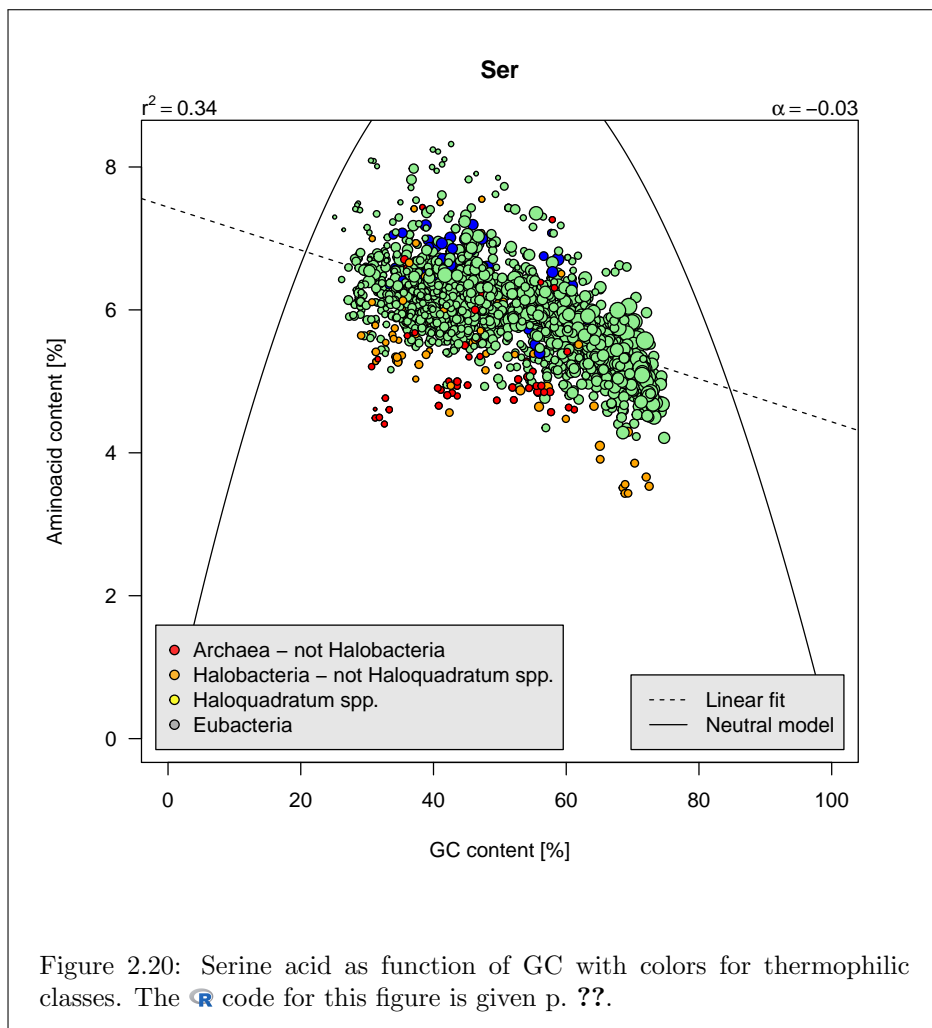
2.4.4 Serine

SERINE is a polar aminoacid encoded by 6 codons. Its frequency ranges on average from 6.7% in low-GC bacteria to 5.2% in high-GC bacteria, that is a factor 1.3. Figure 2.21 page 38 shows that the results are consistent with [136]. The linear model summarises poorly the general trend ($r^2 \approx 0.3$). The bottom outliers are:

```
tdd[tdd$Ser < 4.2, "organism"]
[1] "marinithermus_hydrothermalis" "oceanithermus_profundus"
[3] "rhodothermus_marinus"         "thermaerobacter_marianensis"
[5] "thermaerobacter_subterraneus" "thermus_oshimai"
[7] "thermus_scotoductus"          "thermus_sp"
[9] "thermus_thermophilus"
```

THESE are all thermophilic species but figure 2.20 page 37 show that a low-Ser is not a property shared by all thermophilic species.





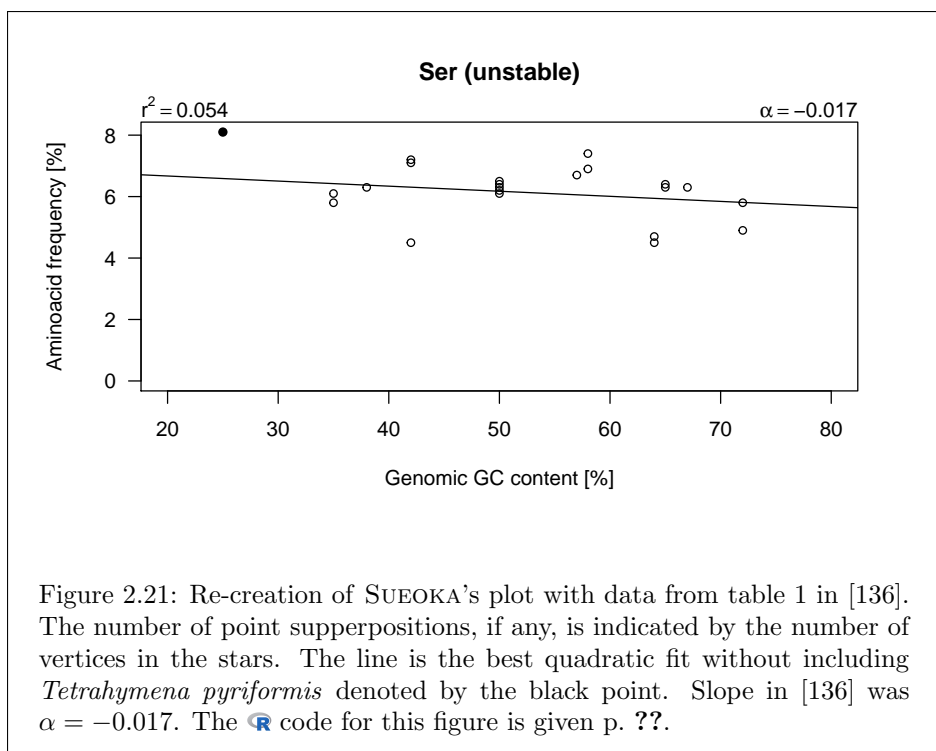

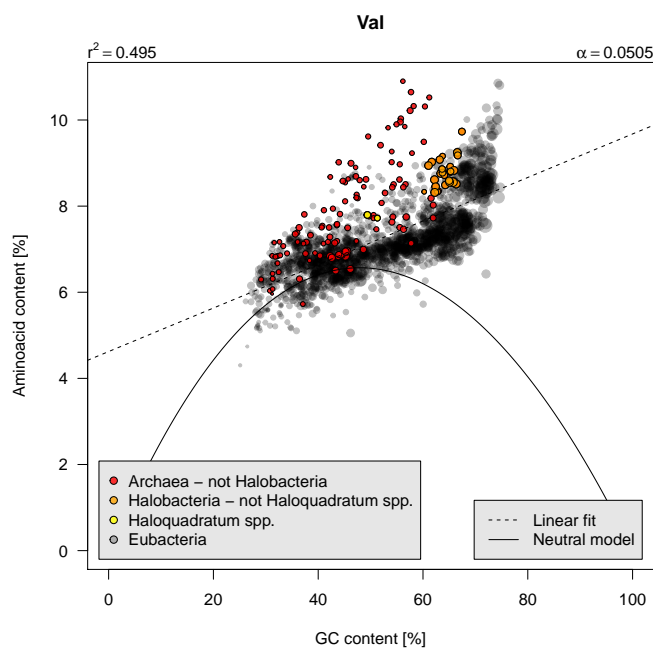
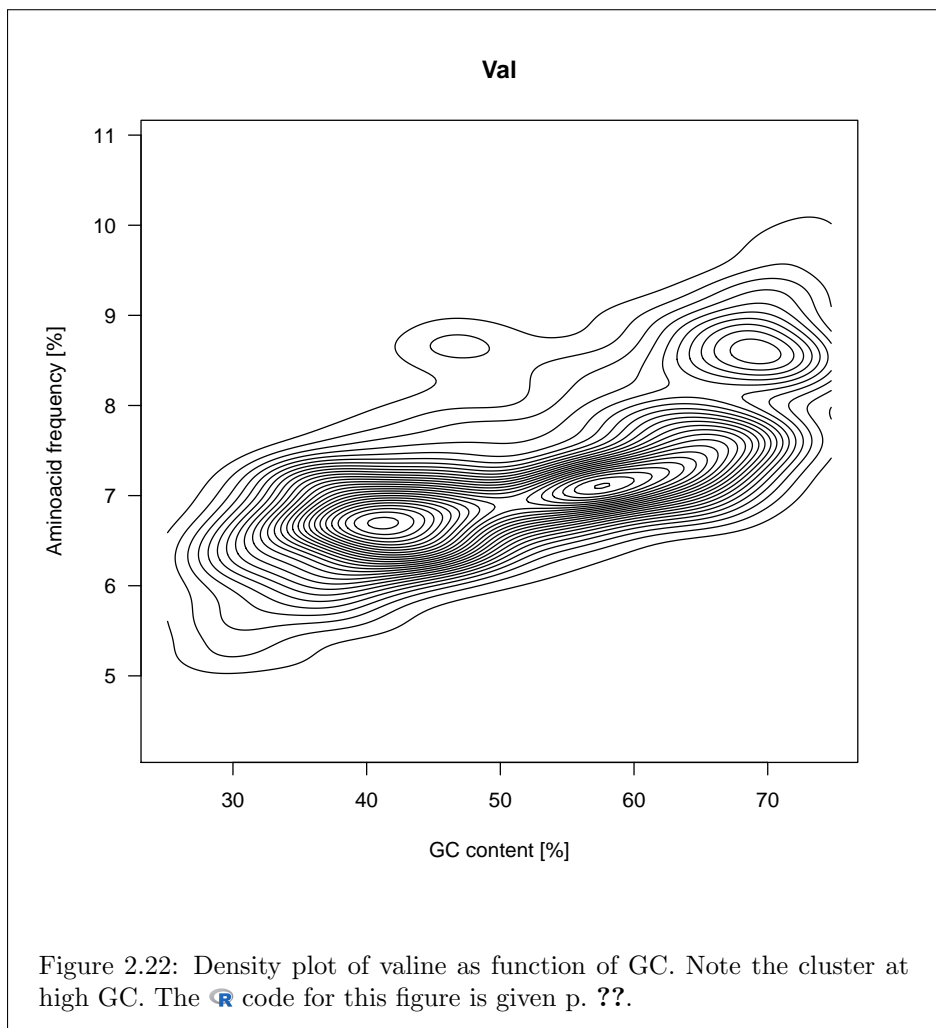


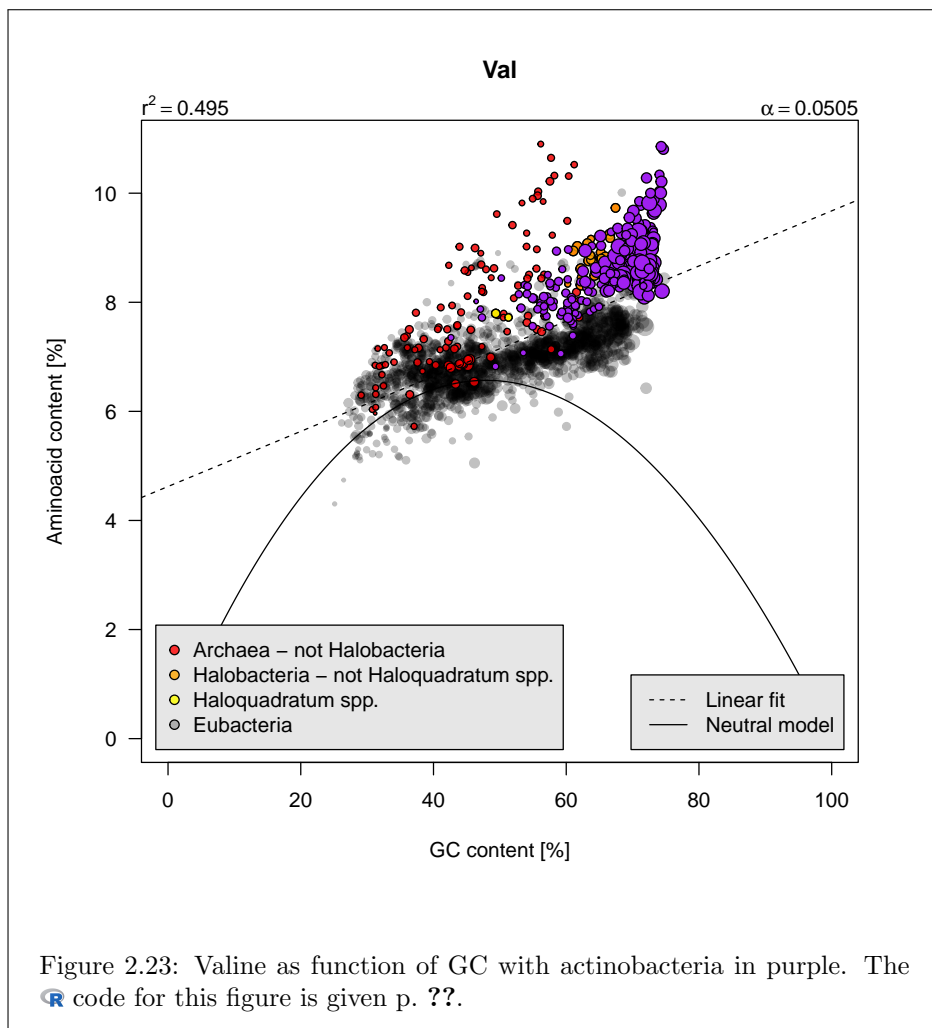
Figure 2.21: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = -0.017$. The  code for this figure is given p. ??.

2.4.5 Valine

VALINE is a non-polar aliphatic aminoacid encoded by 4 codons. Its frequency ranges on average from 6% in low-GC bacteria to 8.5% in high-GC bacteria, that is a factor 1.4. The linear model summarises poorly the general trend ($r^2 \approx 0.5$). Figure 2.24 page 42 shows that this trend was not visible in [136]. Halobacteria tend to favour this aminoacid. There seems to be a cluster of points at high-GC as exemplified by figure 2.22 page 40 and figure 2.23 page 41 shows that it corresponds more or less to the actinobacteria class (high-GC gram+ bacteria TID = 1760).







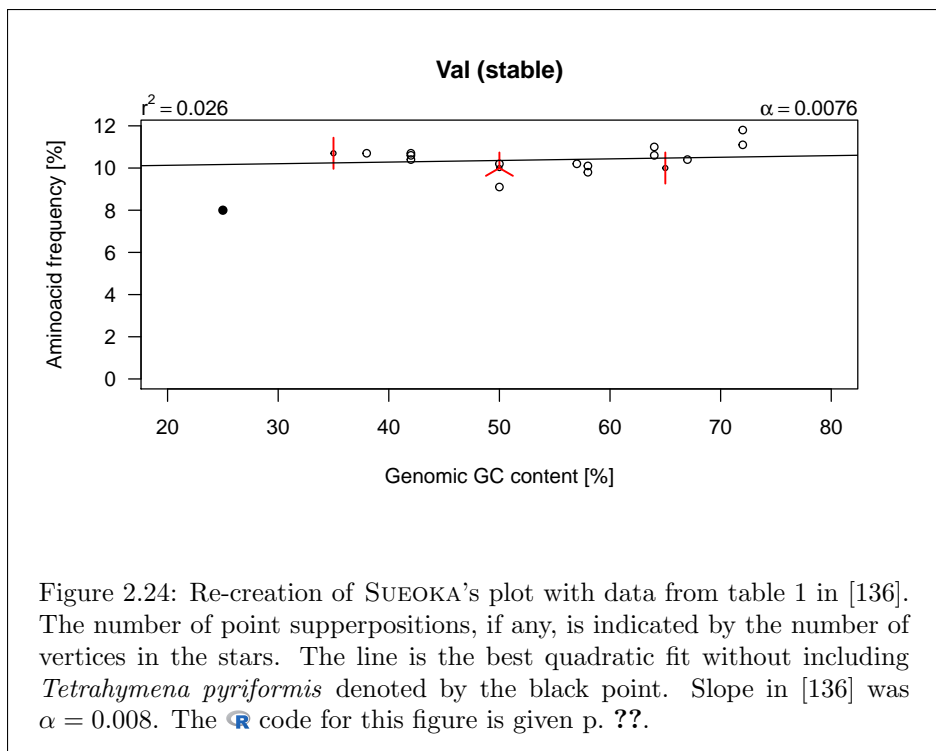
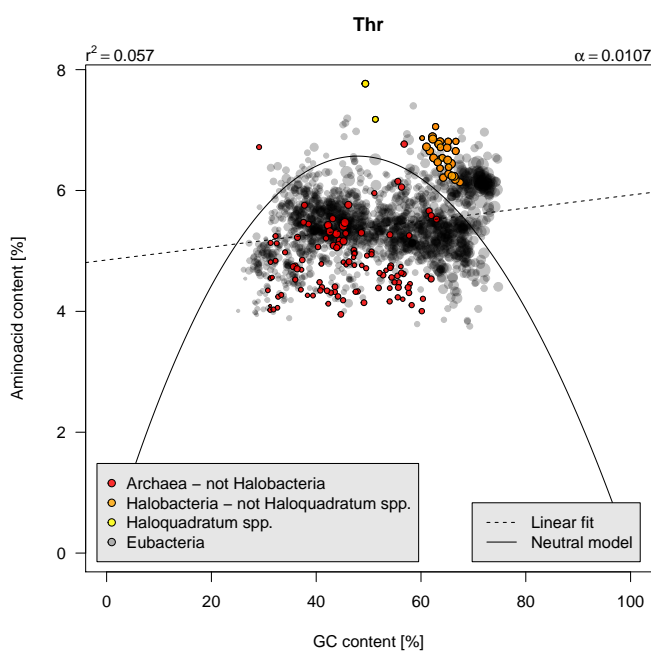


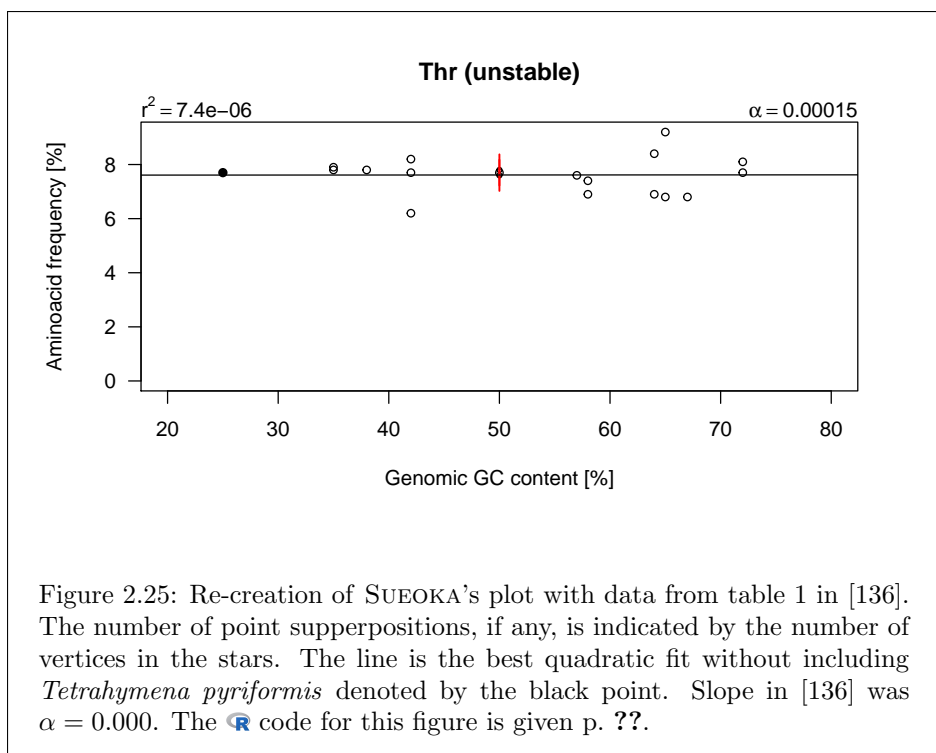
Figure 2.24: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = 0.008$. The `R` code for this figure is given p. ??.

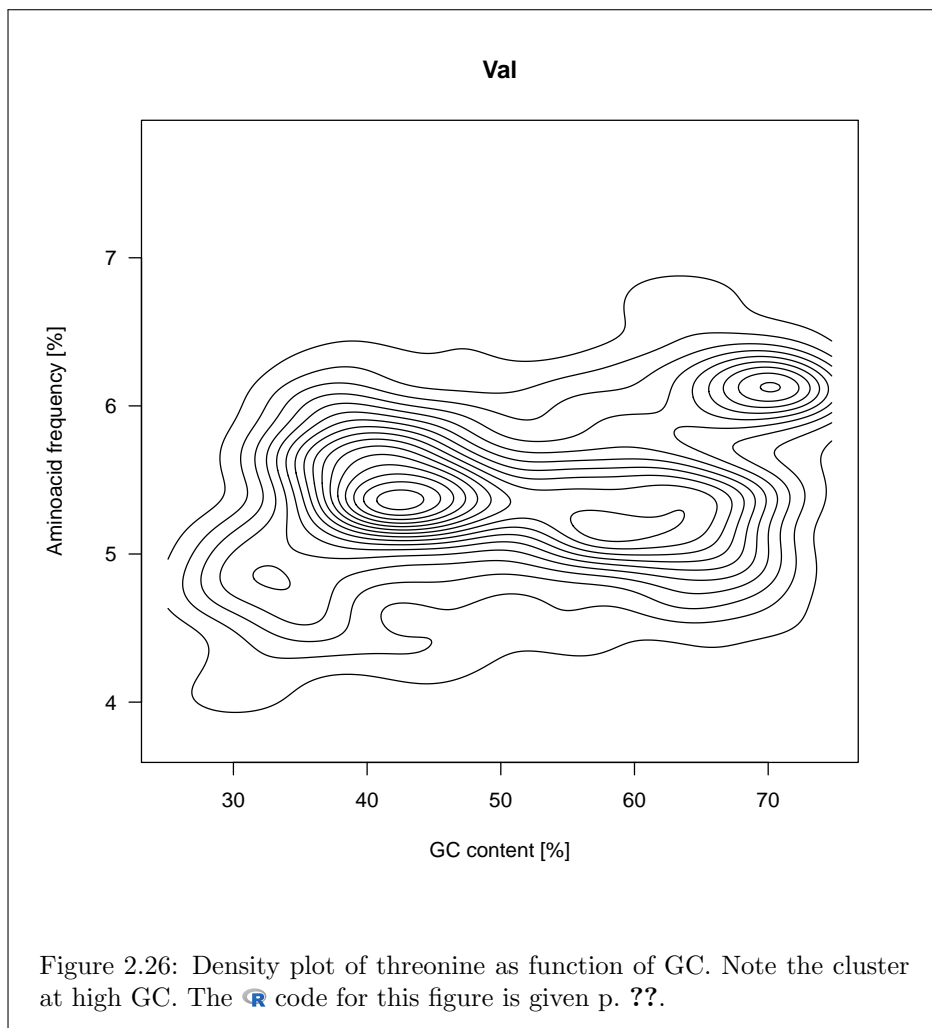
Check bottom outliers

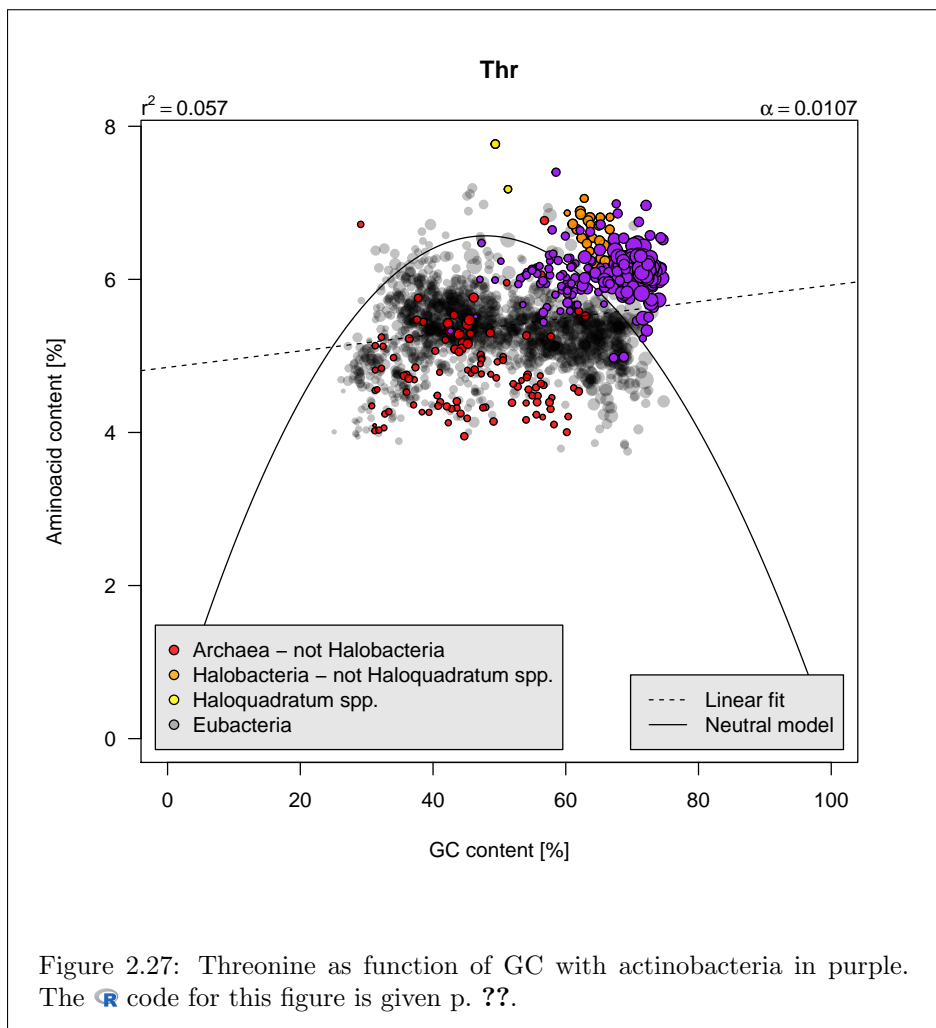
2.4.6 Threonine

THREONINE is a polar, uncharged aminoacid encoded by 4 codons. Its frequency is close to 5% and poorly affected by the GC-content. Figure 2.25 page 44 shows that the results are consistent with [136]. *Haloquadratum walsbyi* is an outlier with a Thr content of 7.8%. As for Val there seems to be a cluster of points at high-GC as exemplified by figure 2.26 page 45 and figure 2.27 page 46 shows that it corresponds more or less to the actinobacteria class (high-GC gram+ bacteria TID = 1760).



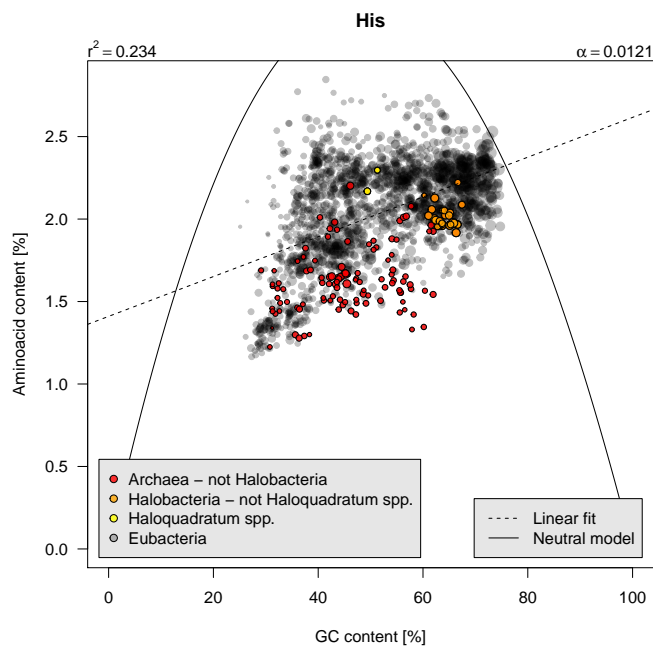


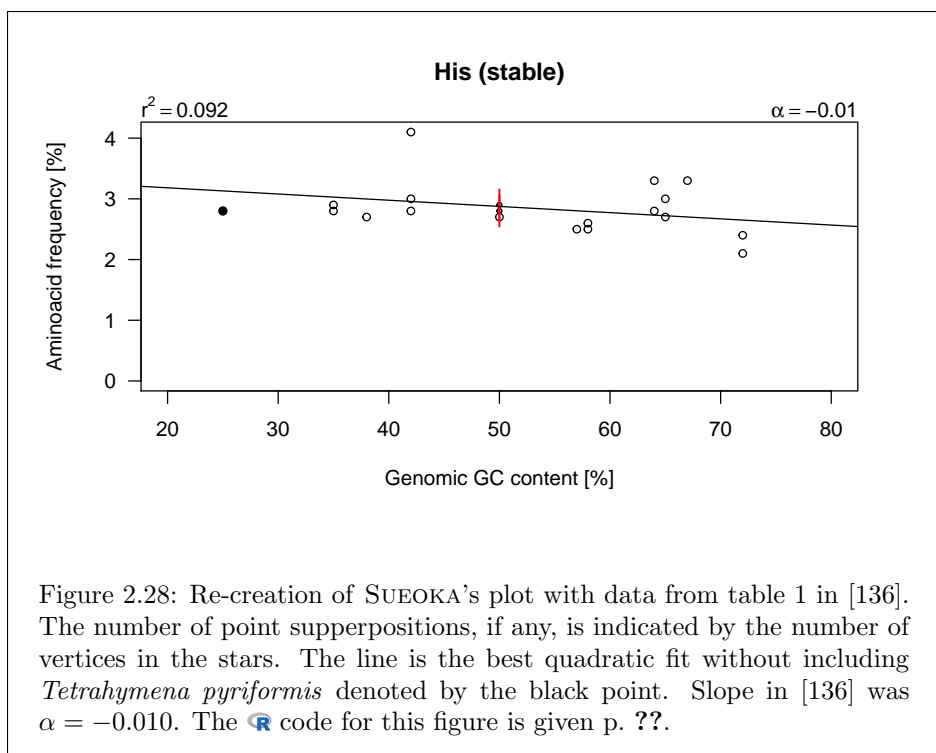




2.4.7 Histidine

HISTIDINE is a positively charged aminoacid encoded by 2 codons. Its frequency ranges on average from 1.7% in low-GC bacteria to 2.3% in high-GC bacteria, that is a factor 1.4. Figure 2.28 page 48 shows that the results are consistent with [136]. The linear model summarises poorly the general trend ($r^2 \approx 0.2$).



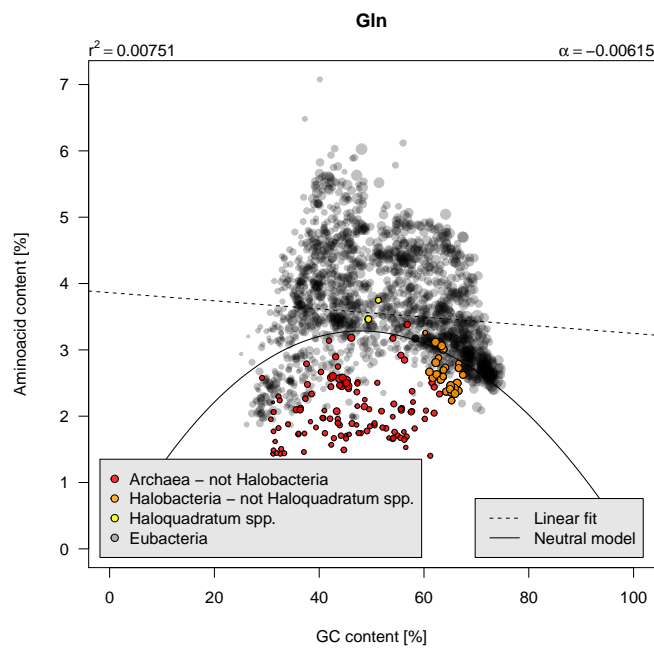


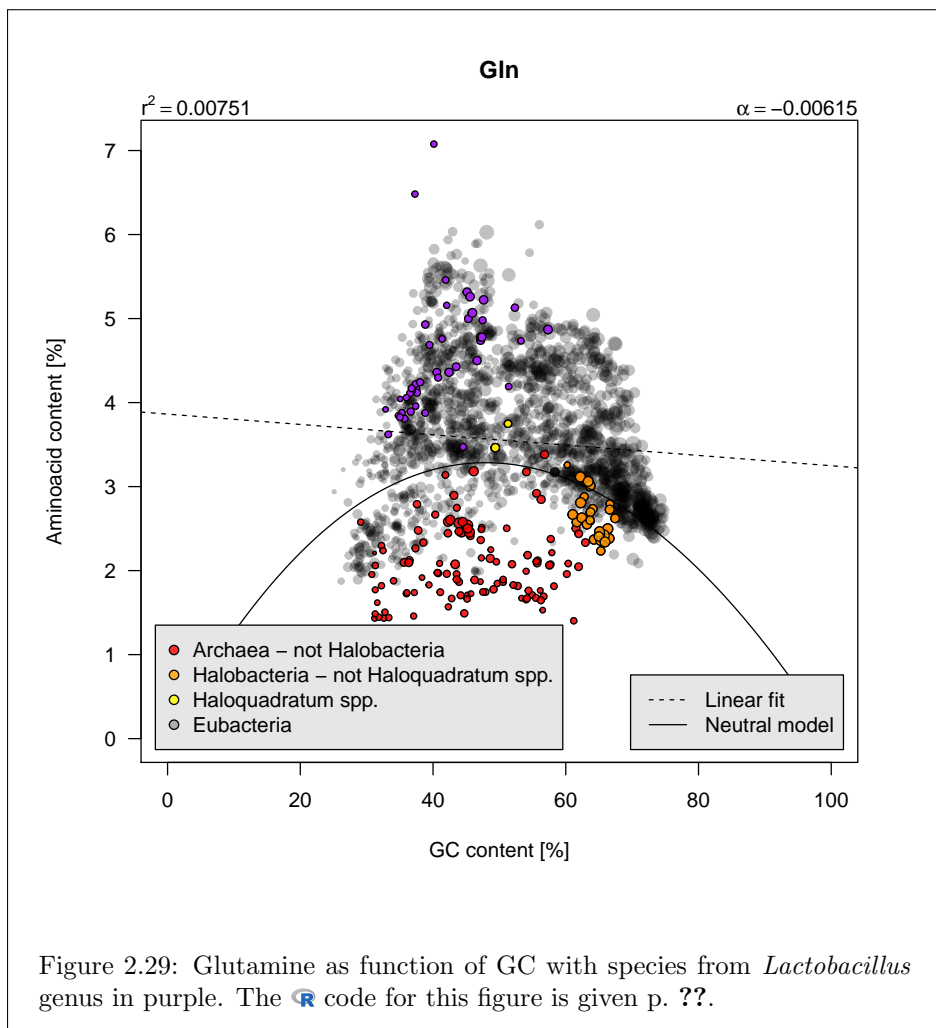
2.4.8 Glutamine

GLUTAMINE is a charge-neutral, polar aminoacid encoded by 2 codons. It is almost unaffected by the GC content with an average frequency of 3.5% close to what would be expected from an uniform codon usage. This aminoacid is clearly avoided in archaea as compared to eubacteria. The top-outliers are:

```
tdd[tdd$Gln > 6.2, "organism"]
[1] "lactobacillus_mellifer" "lactobacillus_mellis"
```

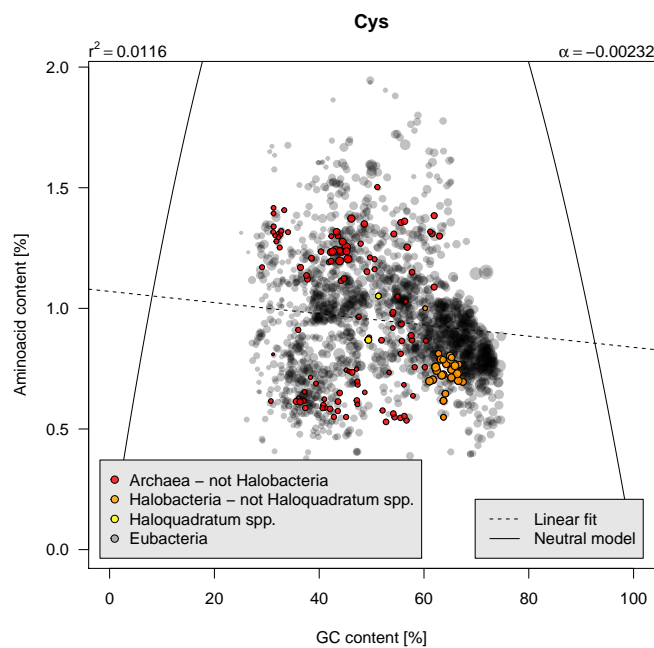
THEY are all from the genus *Lactobacillus* but figure 2.29 page 50 show that this is not a general property of the species from this genus.





2.4.9 Cysteine

CYSTEINE has a thiol side chain and is encoded by 2 codons. Its frequency is poorly affected by the GC content with an average concentration at about 1%, less than what would be expected from a uniform codon usage.

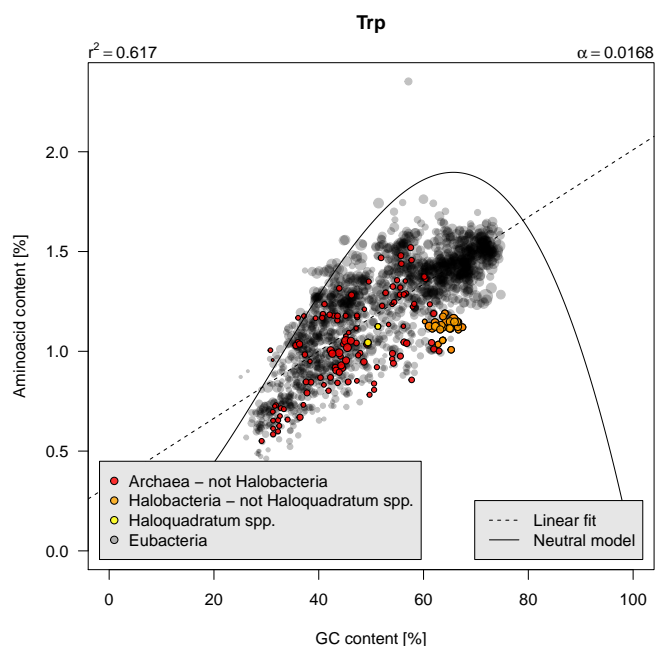


2.4.10 Tryptophane

TRYPTOPHANE is non-polar aromatic amino acid encoded by a single codon. It is a rare amino acid with frequency ranging on average from 0.75% in low-GC bacteria to 1.6% in high-GC bacteria, that is a factor 2. There is a trend for halobacteria to avoid this amino acid. The top outlier is:

```
tdd[tdd$Trp > 2, c("organism", "Trp", "topt")]
          organism      Trp topt
2064 sulfobacillus_acidophilus 2.352197 45
```

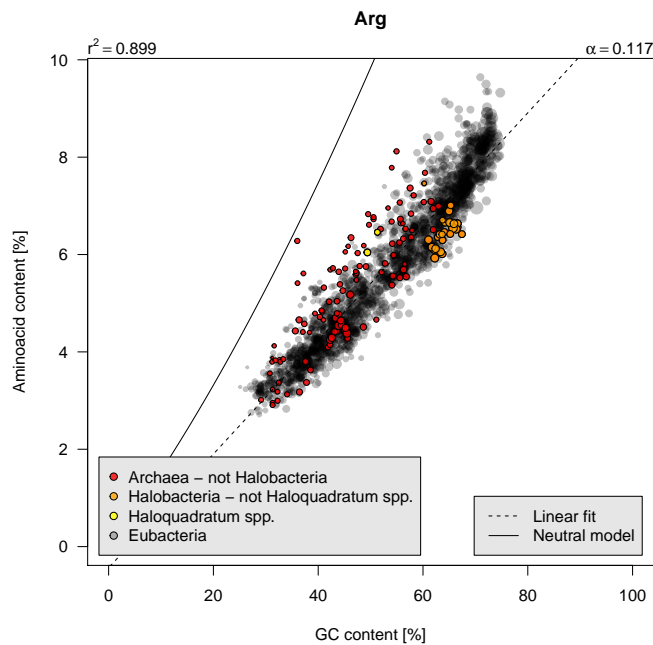
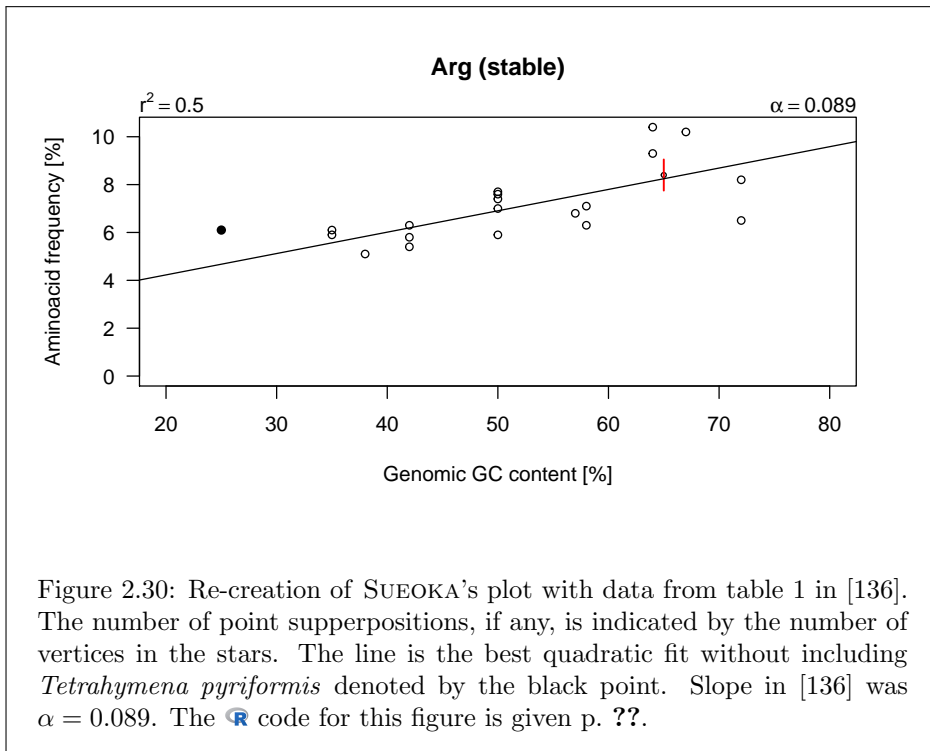
SINCE *Sulfobacillus acidophilus* is the only species available for the *Sulfobacillus* genus here, I can't check if this is a general property.



2.5 Class 3 aminoacids

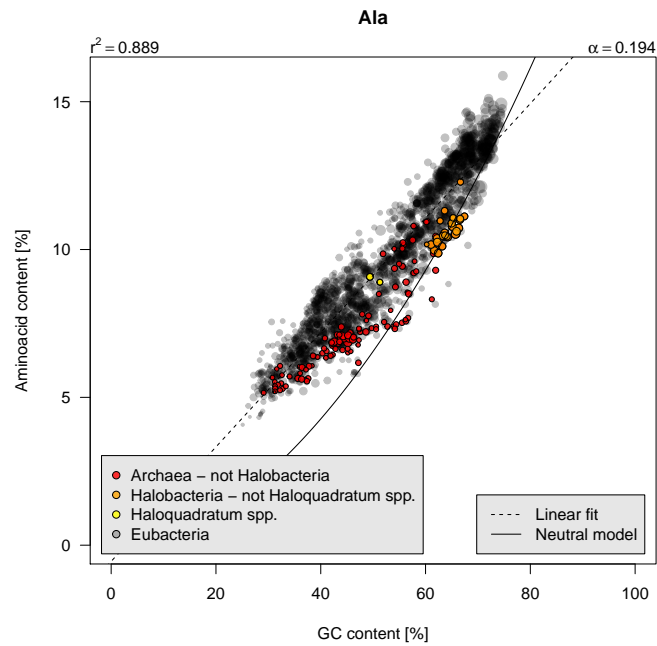
2.5.1 Arginine

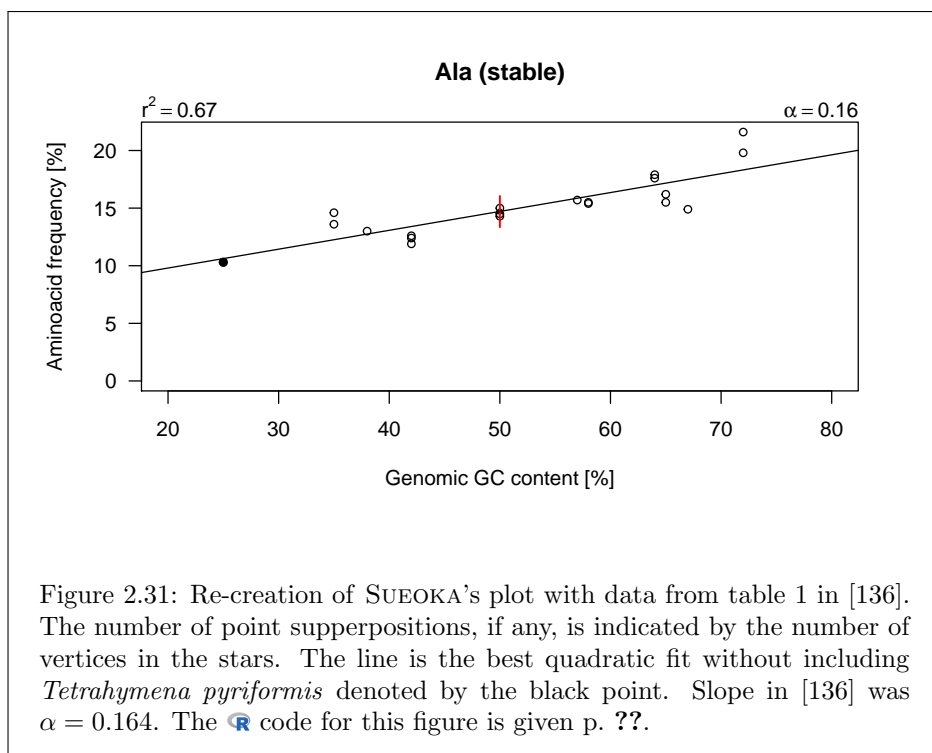
ARGININE is a charged, aliphatic aminoacid encoded by 6 codons. It is very sensitive to the GC content with a frequency ranging on average from 2.6% in low-GC bacteria to 8.3% in high-GC bacteria, that is a factor 3.2. The linear model is a good summary ($r^2 \approx 0.9$). Figure 2.30 page 53 shows that the results are consistent with [136]. All halobacteria but *Haloquadratum walsbyi* tend to avoid this amino-acid.



2.5.2 Alanine

ALANINE is a nonpolar, aliphatic aminoacid encoded by 4 codons. It is very sensitive to the GC content with a frequency ranging on average from 4.5% in low-GC bacteria to 14% in high-GC bacteria, that is a factor 3.1. The linear model is a good summary ($r^2 \approx 0.9$). Figure 2.31 page 55 shows that the results are consistent with [136]. All halobacteria but *Haloquadratum walsbyi* tend to avoid this amino-acid.

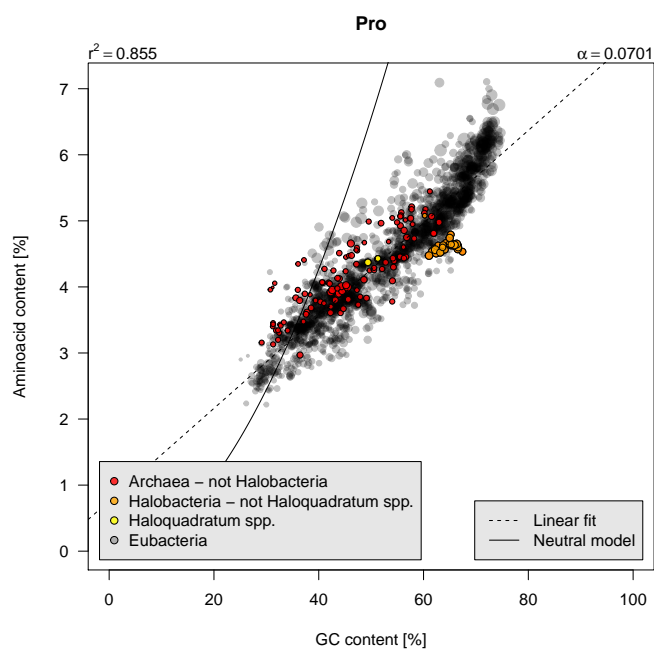


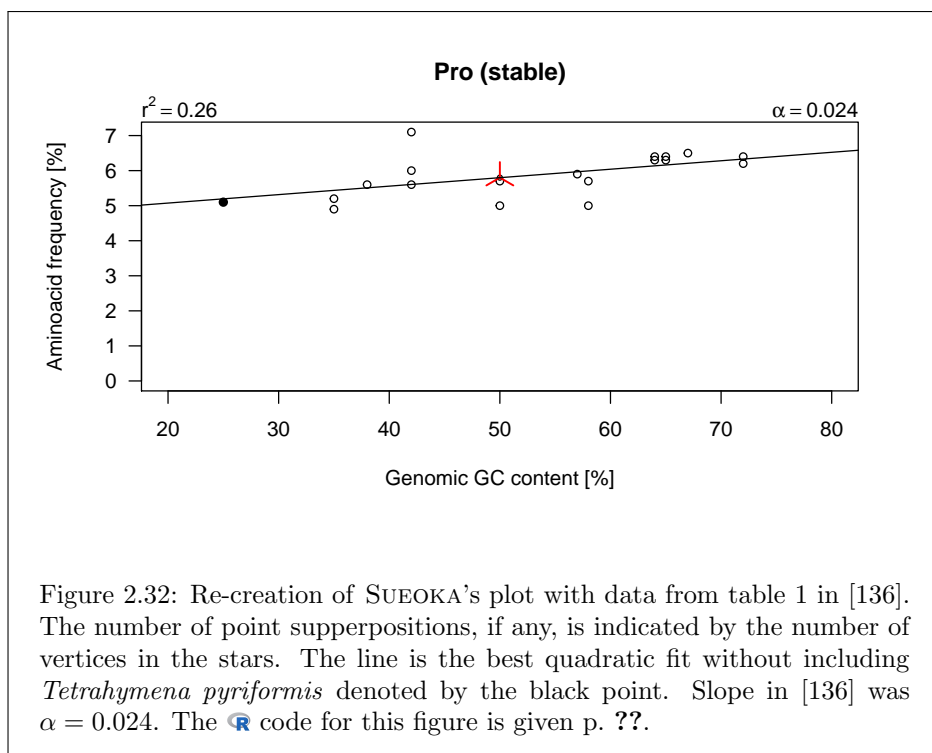


2.5.3 Proline

PROLINE is a nonpolar, aliphatic aminoacid encoded by 4 codons. It is very sensitive to the GC content with a frequency ranging on average from 2.5% in low-GC bacteria to 6% in high-GC bacteria, that is a factor 2.4. The linear model is a good summary ($r^2 \approx 0.9$). Figure 2.32 page 57 shows that the results are consistent with [136]. All halobacteria but *Haloquadratum walsbyi* tend to avoid this amino-acid. The top outliers are:

```
tdd[tdd$Pro > 6.6, "organism"]
[1] "frankia_alni"                "frankia_sp"
[3] "isosphaera_pallida"         "kitasatospora_setae"
[5] "roseomonas_cervicalis"     "streptomyces_cattleya"
[7] "thermaerobacter_marianensis" "thermaerobacter_subterraneus"
```

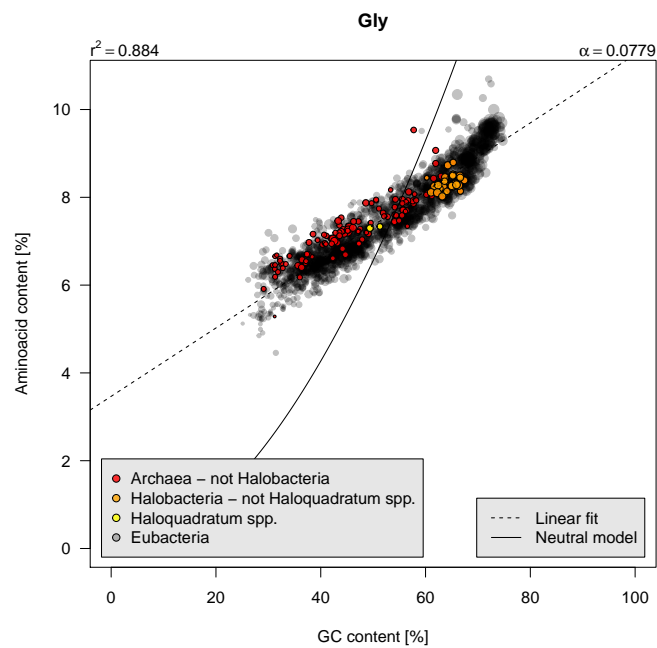


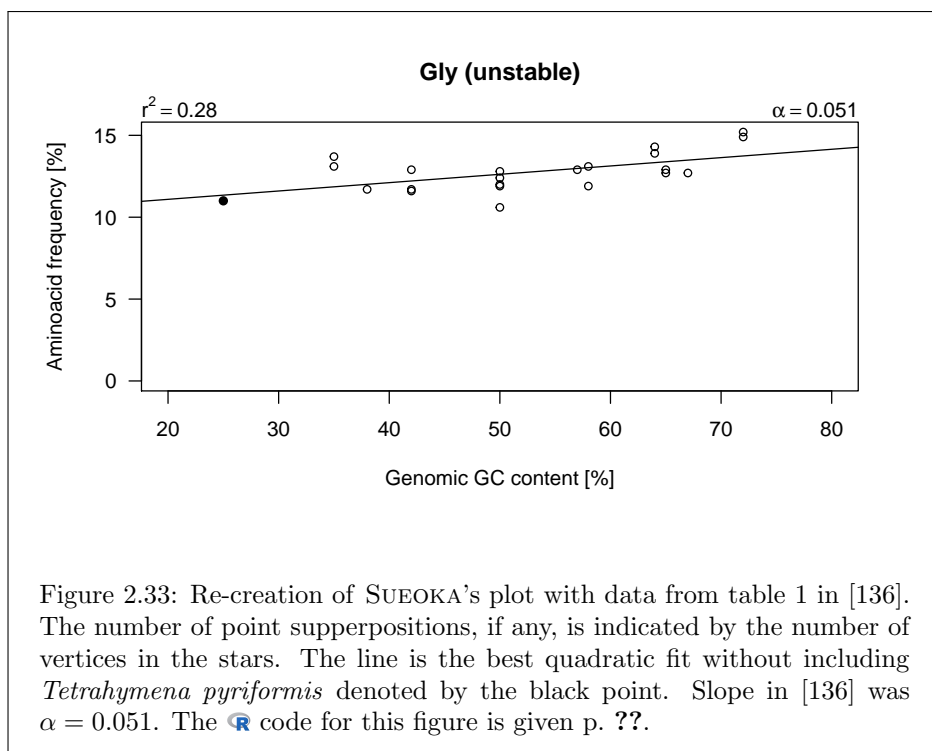


2.5.4 Glycine

GLYCINE has a single hydrogen atom as its side chain and is encoded by 4 codons. It is very sensitive to the GC content with a frequency ranging on average from 5.5% in low-GC bacteria to 9.5% in high-GC bacteria, that is a factor 1.7. The linear model is a good summary ($r^2 \approx 0.9$). Figure 2.33 page 59 shows that the results are consistent with [136]. Top outliers are:

```
tdd[tdd$Gly > 10, "organism"]
[1] "mycobacterium_marinum"      "rubrobacter_xylanophilus"
[3] "thermaerobacter_marianensis" "thermaerobacter_subterraneus"
```





2.6 Evolution of hydrolysis sensitive aminoacids with GC content

2.6.1 Aspartic acid and asparagine

```
showaa(c("Asp", "Asn"))
```

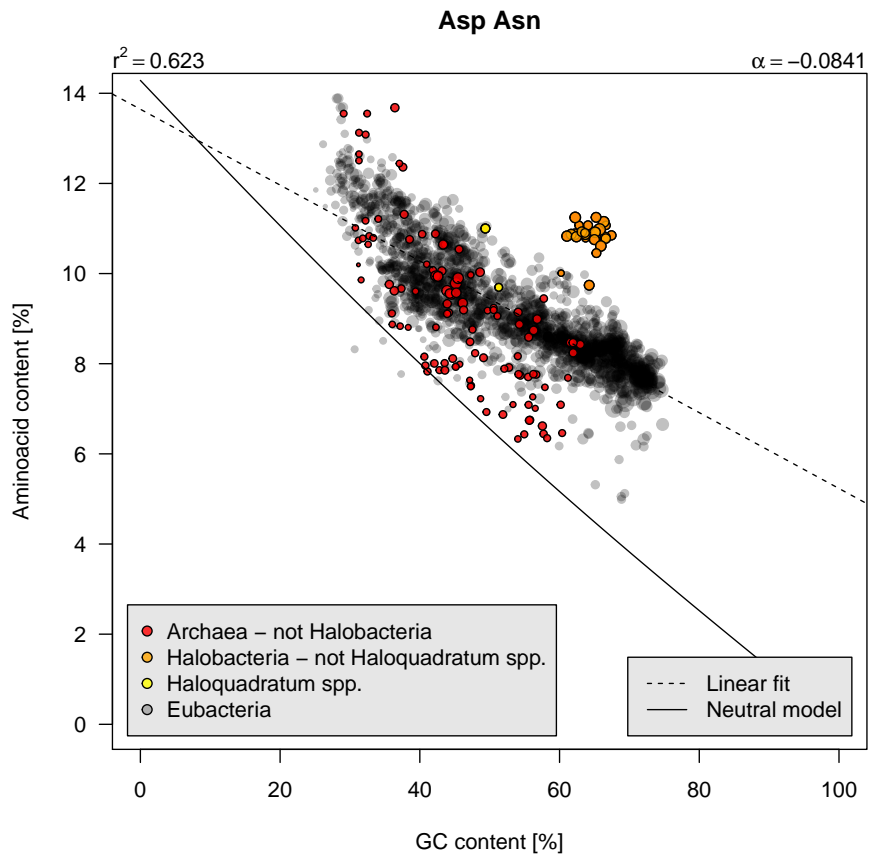


FIGURE 2.34 page 61 shows that the results here are consistent with those from [136].

2.6. EVOLUTION OF HYDROLYSIS SENSITIVE AMINOACIDS WITH GC CONTENT61

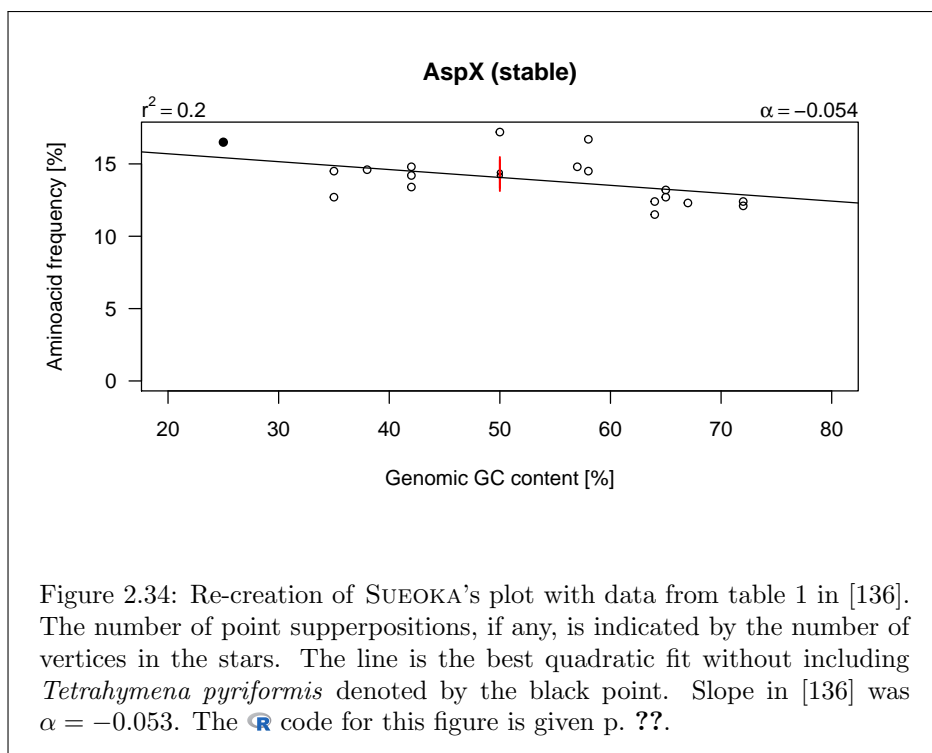


Figure 2.34: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = -0.053$. The R code for this figure is given p. ??.

2.6.2 Glutamic acid and glutamine

```
showaa(c("Glu", "Gln"))
```

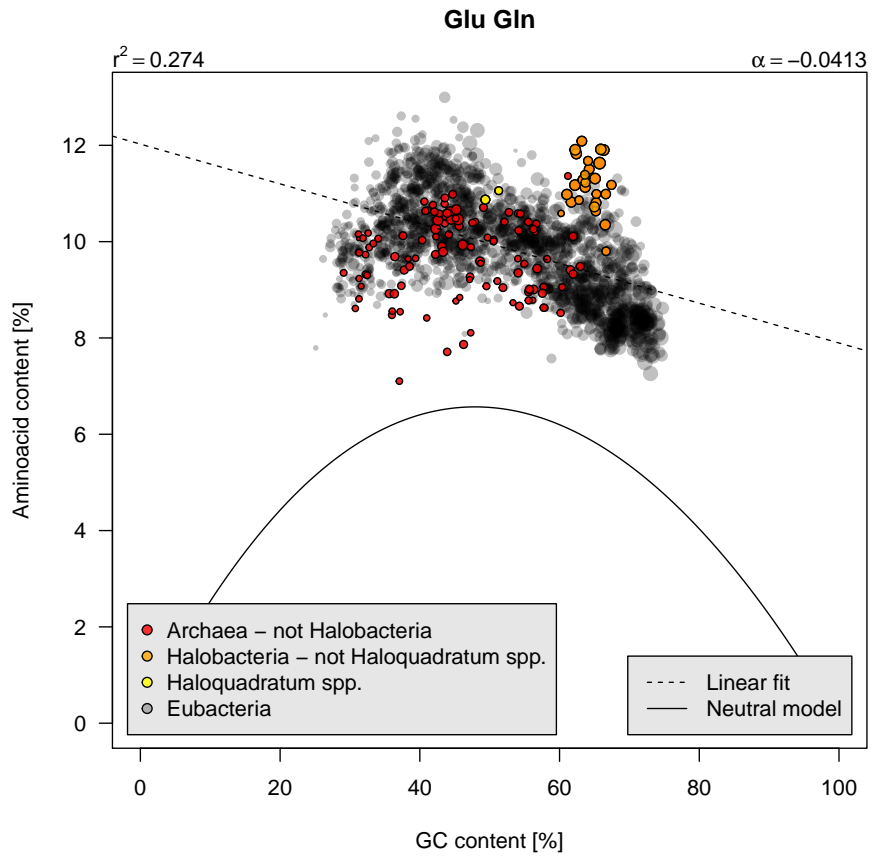


FIGURE 2.35 page 63 shows that the results here are consistent with those from [136].

2.6. EVOLUTION OF HYDROLYSIS SENSITIVE AMINOACIDS WITH GC CONTENT63

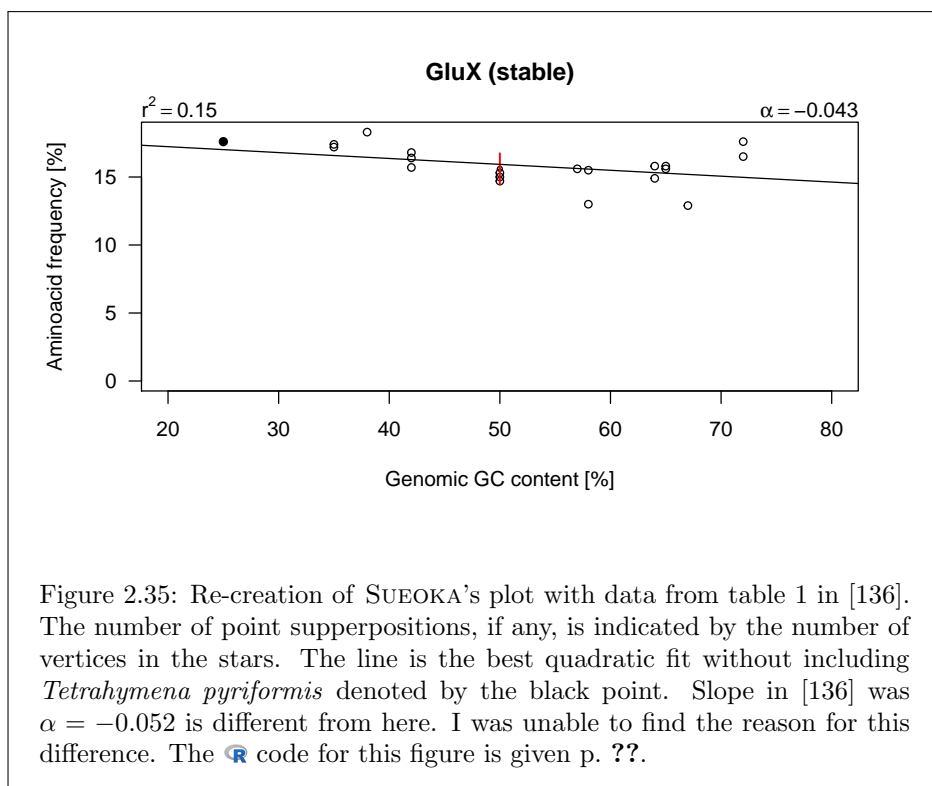


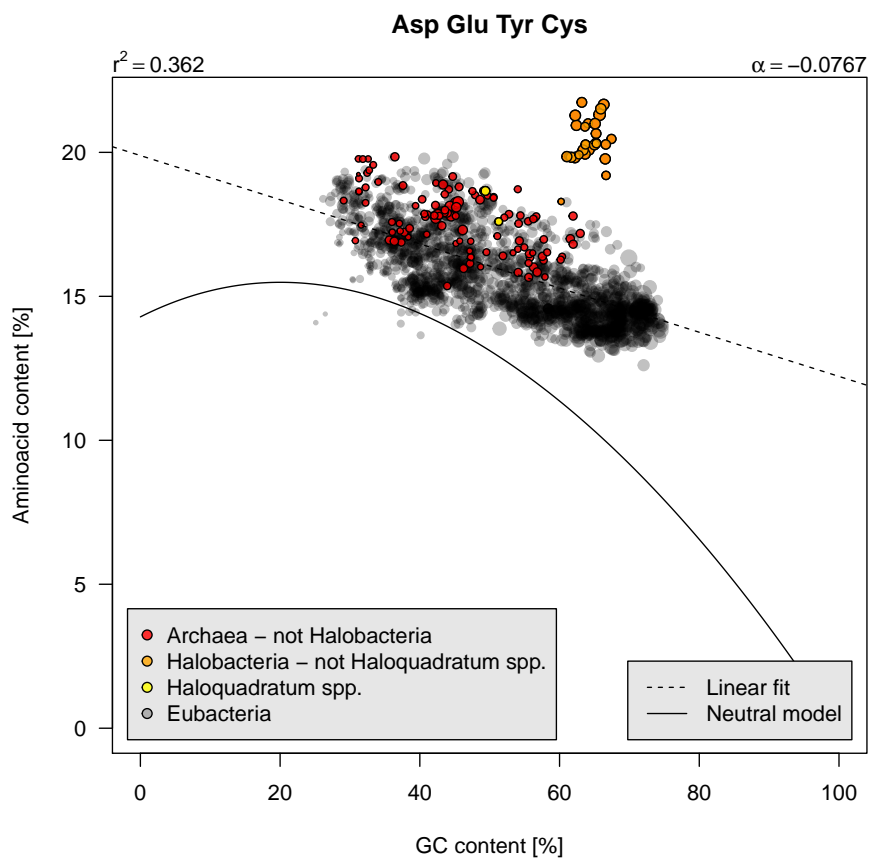
Figure 2.35: Re-creation of SUEOKA's plot with data from table 1 in [136]. The number of point superpositions, if any, is indicated by the number of vertices in the stars. The line is the best quadratic fit without including *Tetrahymena pyriformis* denoted by the black point. Slope in [136] was $\alpha = -0.052$ is different from here. I was unable to find the reason for this difference. The `R` code for this figure is given p. ??.

2.7 Evolution of charged aminoacids with GC content

2.7.1 Negatively charged aminoacid

THE frequency of negatively charged aminoacids decreases with GC content from 18% in low-GC bacteria to 14% in high-GC bacteria, that is a factor 1.3.

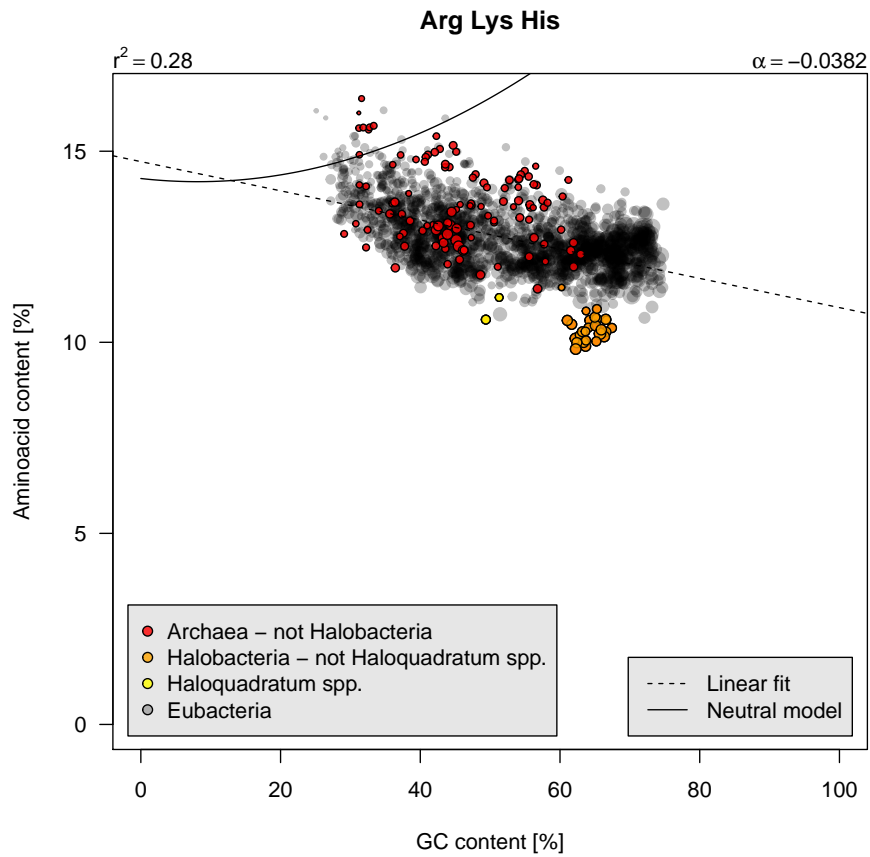
```
showaa(c("Asp", "Glu", "Tyr", "Cys"))
```



2.7.2 Positively charged aminoacid

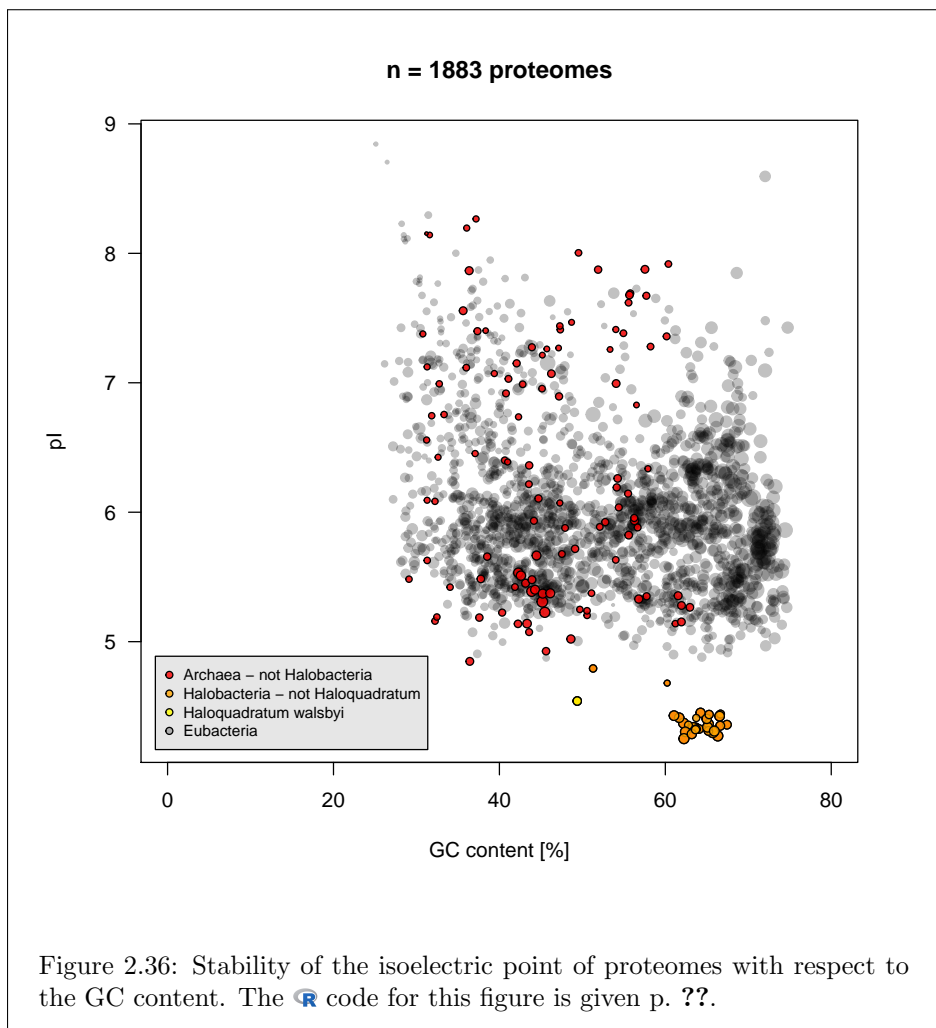
THE frequency of positively charged aminoacids also decreases with GC content from 13.8% in low-GC bacteria to 11.8% in high-GC bacteria, that is a factor 1.2. Note that the frequency of positively charged aminoacids is on the opposite expected to increase with GC content, we have perhaps here a selective pressure on the global charge of the proteins.

```
showaa(c("Arg", "Lys", "His"))
```

2.7.3 Evolution of pI with GC

As we may have expected from the simultaneous decrease of positively and negatively charged aminoacids described in the two previous sections, figure 2.36 page 66 shows that pI is unaffected by the GC content.



2.8 Summary of outstanding bacterial groups

It long has been recognized that proteins from “extreme halophiles” are acidic, see for instance the 1974 review [66].

Summarize here what was gained from univariate analysis of aminoacid usage in bacteria

Chapter 3

Multivariate analysis of aminoacid usage

3.1 Loading the dataset

```
load("local/tdd.Rda")
```

3.2 Utilities

3.2.1 First factorial map orientation

3.3 Sanity check

IN this section I want to check that the results are the same when computing I correspondence analysis of aminoacid usage in two different ways.

Automate axis orientation so as to have always low-GC on the left and thermophilic on the top

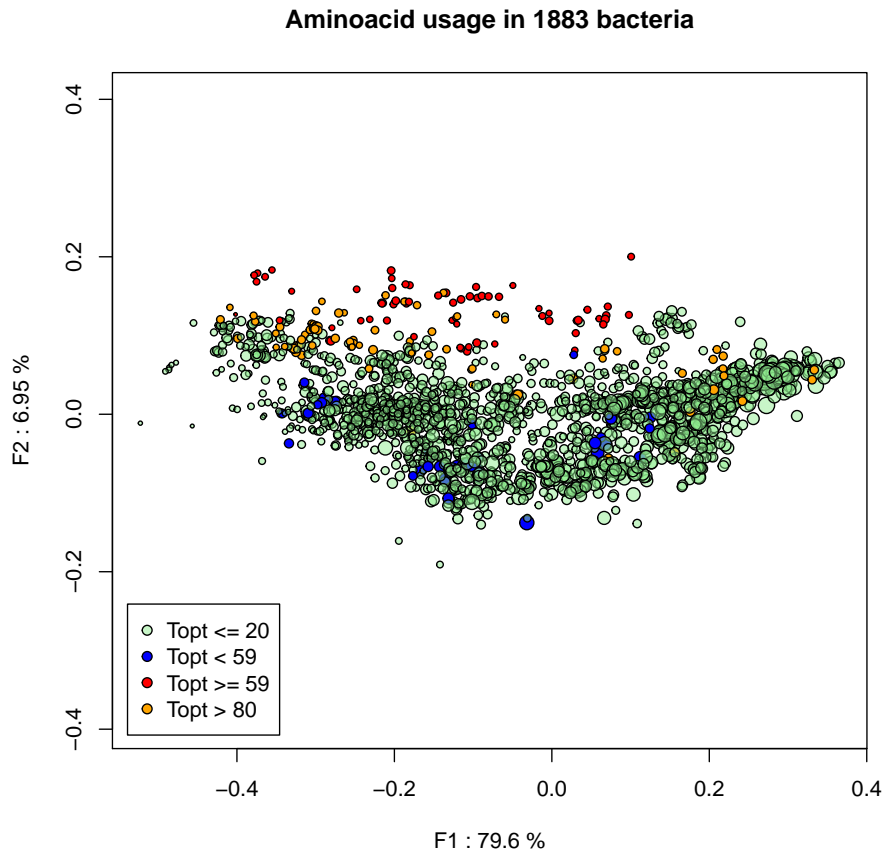
3.3.1 Direct CA on aminoacid frequencies

```
codons <- colnames(tdd[, 2:65])
facaa <- factor(sapply(codons, function(x) aaa(translate(s2c(x)))))
tdaa <- t(apply(tdd[, 2:65], 1, function(x) tapply(x, facaa, sum)))
library(ade4)
checkcoa1 <- dudi.coa(tdaa, scannf = FALSE, nf = 2)
swap <- function(dudi, nf){
  dudi$li[, nf] <- -1*dudi$li[, nf]
  dudi$co[, nf] <- -1*dudi$co[, nf]
  return(dudi)
}
checkcoa1 <- swap(checkcoa1, 1) # High GC on right
checkcoa1 <- swap(checkcoa1, 2) # Thermophiles on top
checkplot <- function(dudi){
  main <- paste("Aminoacid usage in",
               nrow(dudi$tab), "bacteria")
  plot(dudi$li[,1], dudi$li[,2], pch = 21, bg = tdd$athermocol,
       asp = 1, cex = tdd$cex, xlab = "", ylab = "", main = main)
  title(xlab = paste("F1 :", signif(100*dudi$eig[1]/sum(dudi$eig), 3), "%"))
  title(ylab = paste("F2 :", signif(100*dudi$eig[2]/sum(dudi$eig), 3), "%"))
}
```

```

legend("bottomleft", inset = 0.02,
      legend = c("Topt <= 20", "Topt < 59", "Topt >= 59", "Topt > 80"),
      pch = 21, pt.bg = unique(tdd$a_thermocols))
}
checkplot(checkcoa1)

```

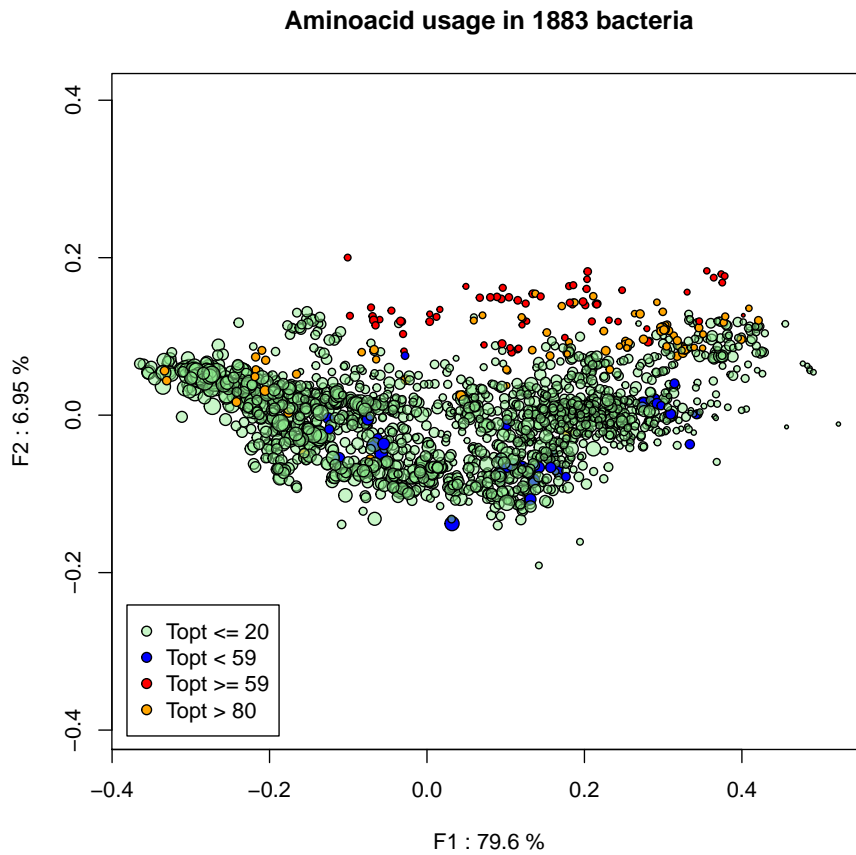


3.3.2 BCA on codon frequencies

```

library(seqinr)
tdco <- tdd[, 2:65]
codons <- colnames(tdco)
faca <- factor(sapply(codons, function(x) aaa(translate(s2c(x)))))
coa <- dudi.coa(tdco, scannf = F, nf = 2)
checkcoa2 <- t(bca(t(coa), faca, scannf = FALSE, nf = 2))
checkcoa2 <- swap(checkcoa2, 2) # Thermophiles on top
checkplot(checkcoa2)

```



3.3.3 Comparisons

```

all.equal(checkcoa1$eig, checkcoa2$eig)
[1] TRUE
all.equal(checkcoa1$li[, 1], checkcoa2$li[, 1])
[1] "Mean relative difference: 2"
all.equal(checkcoa1$li[, 2], checkcoa2$li[, 2])
[1] TRUE

```

3.3.4 Conclusion

As expected, the results are exactly the same. This is because CA on amino acid frequencies is the same analysis as the between group analysis for CA on codon frequencies.

Chapter 4

Univariate analysis of synonymous codon usage

4.1 Utilities definition

4.1.1 Loading the dataset

```
load("local/tdd.Rda")
```

4.1.2 Computing codon relative frequencies

```
library(seqinr)
codons <- colnames(tdd[, 2:65])
facao <- factor(sapply(codons, function(x) aaa(translate(s2c(x)))))
tdc <- t(apply(tdd[, 2:65], 1, function(x)
  unlist(tapply(x, facaa, function(y) 100*y/sum(y)))))
tdc <- as.data.frame(tdc)
substr(names(tdc), 4,4) <- "-" # remove the dot for file names
tdc$tdgc <- tdd$tdgc
```

4.1.3 Plotting data

```
plotcod <- function(codlist){
  x <- tdc$tdgc
  y <- rowSums(cbind(tdc[, which(colnames(tdc) %in% codlist)], 0))
  plot(x, y, xlim = c(0, 100), ylim = c(0, 100), las = 1,
       xlab = "GC content [%]", ylab = "Codon relative frequency [%]",
       pch = 19, cex = tdd$cex, main = paste(codlist, collapse = " "), col = col2alpha("black", 0.25))
  abline(lm(y~x), lty = 2)
  abline(v = 50, lty = 2)
  abline(h = mean(y), lty = 2)
  axis(4)
  mtext(bquote(r^2 == .(signif(cor(x,y)^2, 3))), adj = 0.5)
  mtext(paste("Slope =", signif(lm(y~x)$coef[2], 3)), adj = 0)
  mtext(paste("Intercept =", signif(lm(y~x)$coef[1], 3)), adj = 1)
  isa <- which(tdd$domain == "Archaea")
  points(x[isa], y[isa], pch = 21, bg = col2alpha("red", 0.8), cex = tdd$cex[isa])
  ish <- which(tdd$class == 183963) # Halobacteria
  points(x[ish], y[ish], pch = 21, bg = col2alpha("orange", 0.8), cex = tdd$cex[ish])
  lines(lowess(x, y), col = "red")
  ou <- ifelse(lm(y~x)$coef[2] > 0, "topleft", "topright")
```

```

legend(ou, inset = 0.02, legend = c("Archaea - not Halobacteria",
  "Halobacteria", "Eubacteria"), pch = 21,
  pt.bg = c(col2alpha("red", 0.8), col2alpha("orange", 0.8),
  col2alpha("black", 0.25)),
  bg = grey(0.9))
}

```

4.1.4 Generation of all figures

This code is used to generate all figures:

```

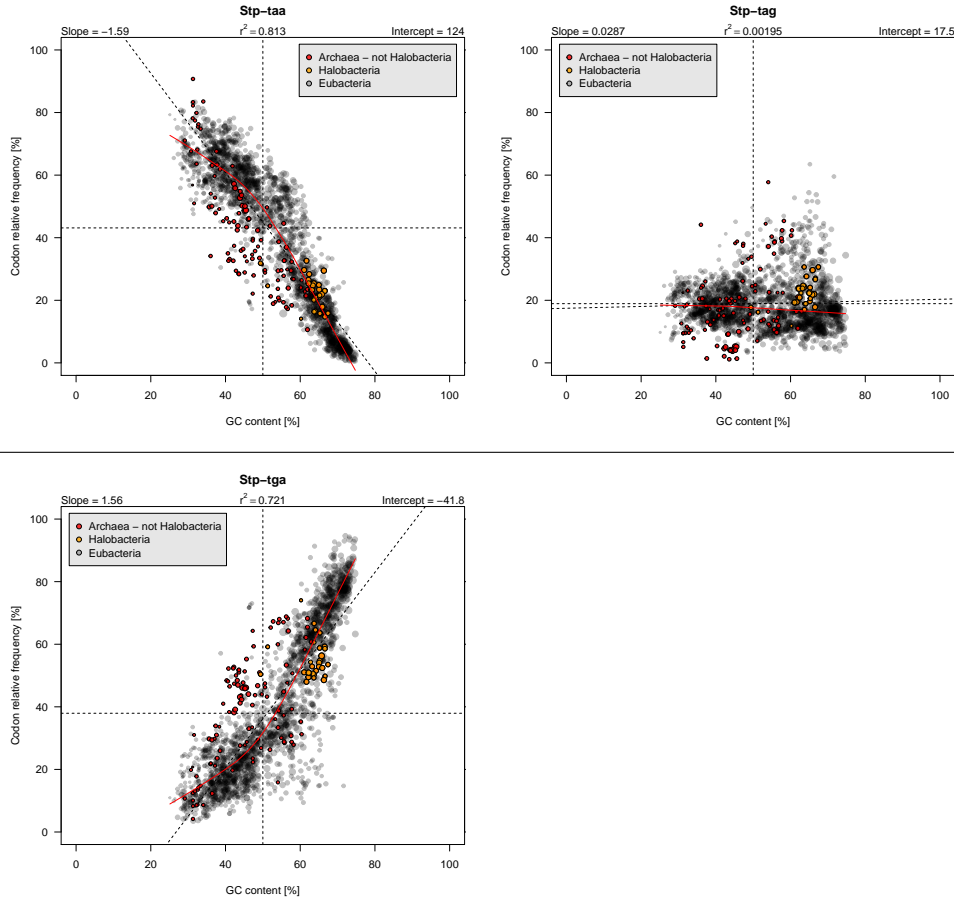
todo <- names(tdc)[-65]
for(i in todo){
  fname <- paste("figs/auto-", i, ".pdf", sep = "")
  pdf(fname)
  par(mar = c(5, 4, 4, 0) + 0.1)
  plotcod(i)
  dev.off()
}

```

4.2 Introduction

BECAUSE there are 64 codons, we need some guideline to structure this chapter. I will follow here an old typology given by figure 1 in [41].

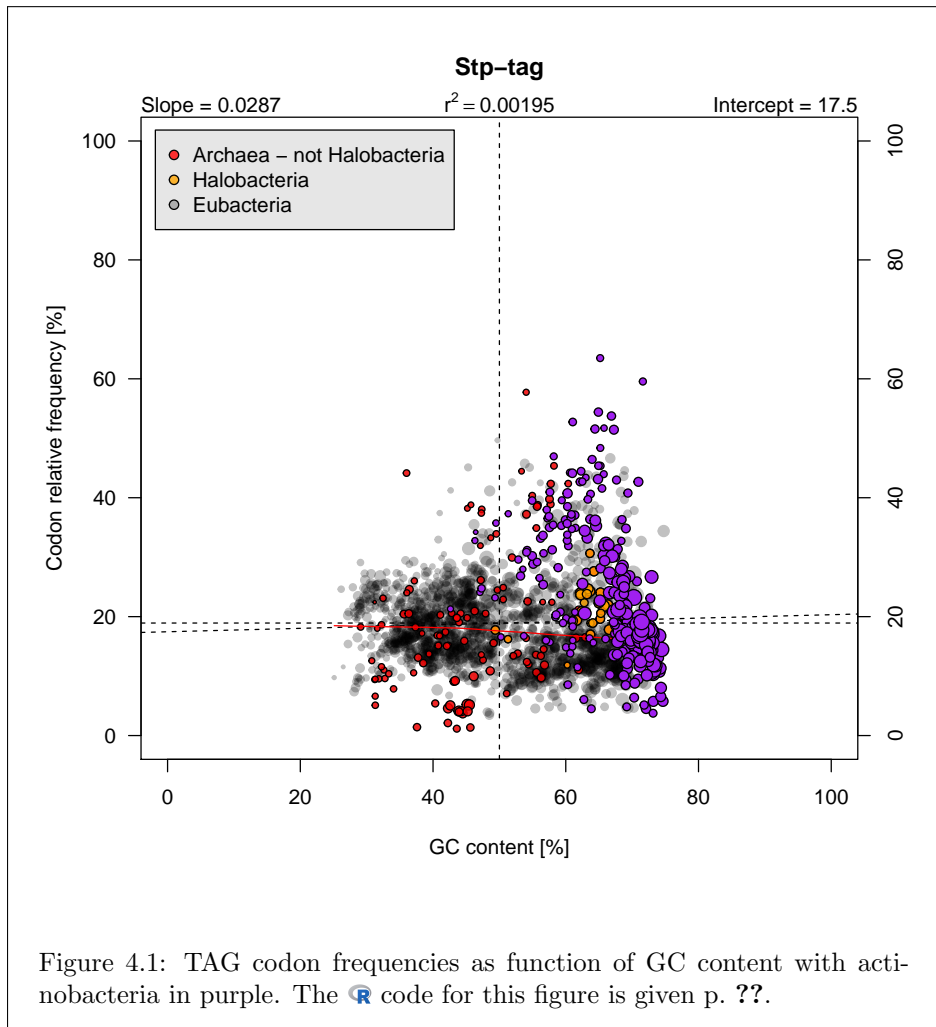
4.3 Terminators



THERE is a strong influence of the GC content on TAA and TGA but TAG is almost unaffected and avoided, consistently with previous results [108, 61]. In most bacteria, TAA and TGA represent about 80% of stop codons with TAA favoured in low GC bacteria and TGA favoured in high GC bacteria. TAG is recognized only by RF1, TGA is recognized only by RF2, and TAA is recognized by both factors [125] and is the major codon in genes with a high expressivity [61]. TAG is usually avoided¹, but there are some outliers with more than 50% of this codon:

```
tdd[tdc$`Stp-tag` > 50, c("organism", "phylum", "class", "order", "family")]
      organism phylum class order family
92  adlercreutzia_equolifaciens 201174 84998 1643822 1643826
617 collinsella_tanakaei 201174 84998 84999 84107
637 corynebacterium_atypicum 201174 1760 85007 1653
673 corynebacterium_vitaeruminis 201174 1760 85007 1653
842 enterorhabdus_caecimuris 201174 84998 1643822 1643826
1056 hyperthermus_butylicus 28889 183924 114380 2307
1490 olsenella_uli 201174 84998 84999 1643824
1833 rubrobacter_radiotolerans 201174 84995 84996 84997
2221 turicella_otitidis 201174 1760 85007 1653
```

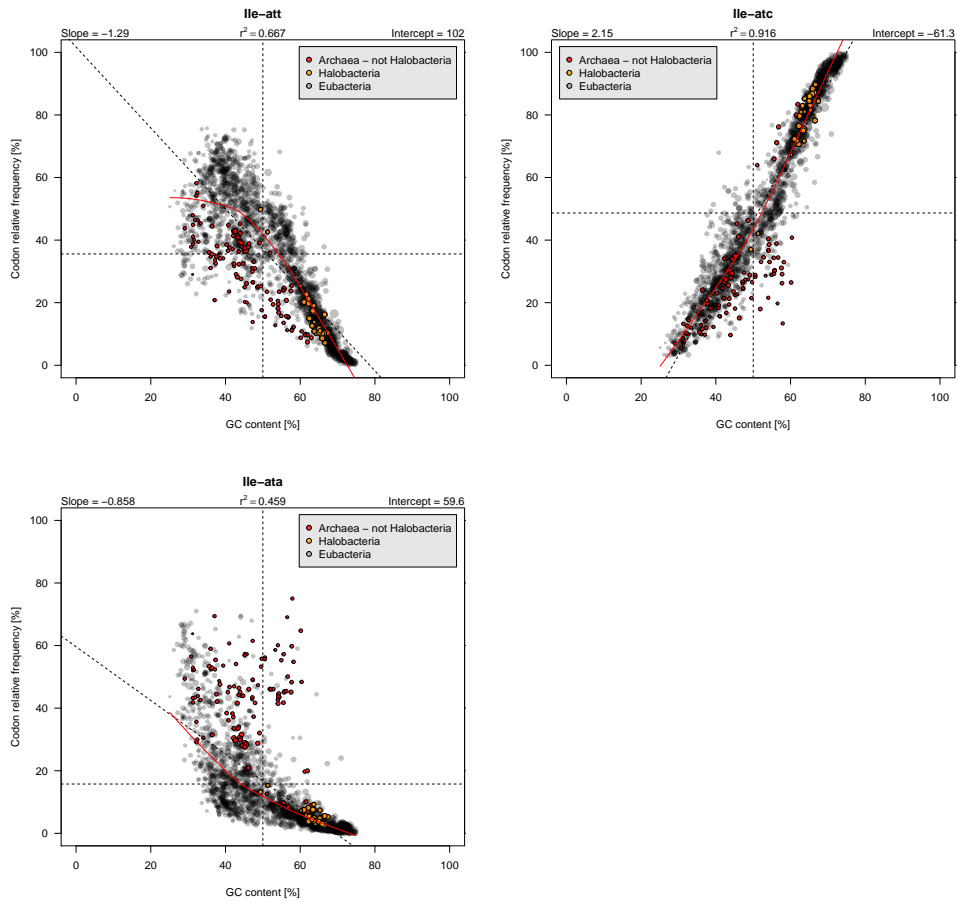
¹In *Escherichia coli*, *Mycobacterium smegmatis* and *Bacillus subtilis* the TAG/TGA frequency ratio matches well with the RF1/RF2 intracellular concentration ratio [61].



MOST outliers are actinobacteria (TID 201174). However, figure 4.1 page 76 shows that this is not a characteristic shared by all members of this group.

4.4 Odd number

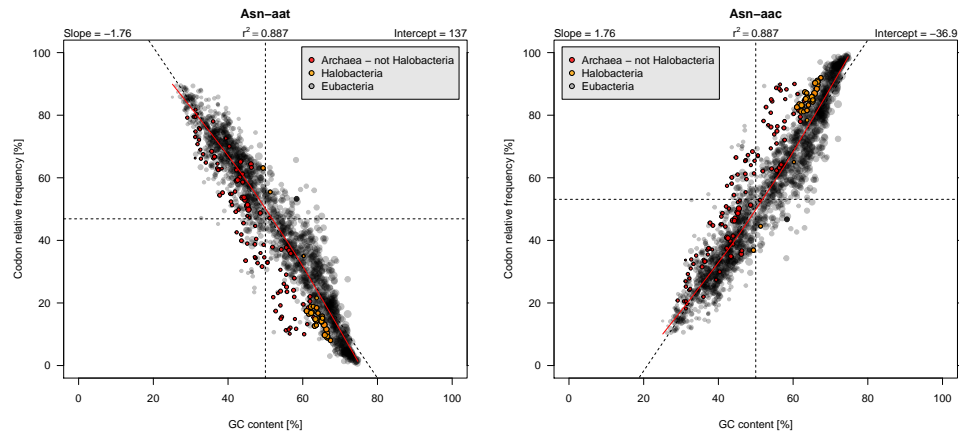
THERE is a single codon for Met and Trp so that their relative frequencies are always 100%, which is uninformative. The only odd number interesting class is for Ile.



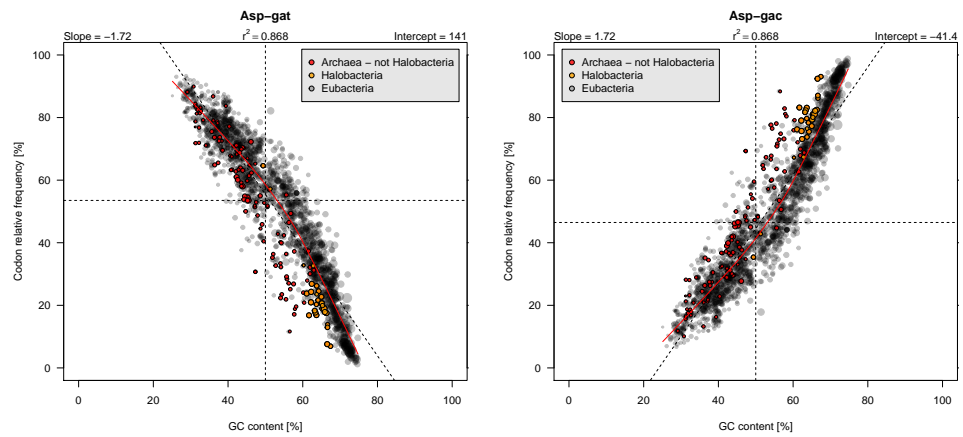
FOR leucine the order of preference for codons is $ATC > ATT > ATA$. There is much variability at low GC than high GC because at high GC ATC is almost exclusive while at low GC there is freedom left between ATT and ATA. As a consequence the linear model is very good for ATC and bad for ATT and ATA since we have a triangular relationship rather than a linear one. The linear relationship would be of course restored by summing ATT and ATA since this yields a relationship symmetrical the ATC one. Archaea but halobacteria are depleted in ATT and ATC and then enriched in ATA codons as compared to eubacteria.

4.5 Duet

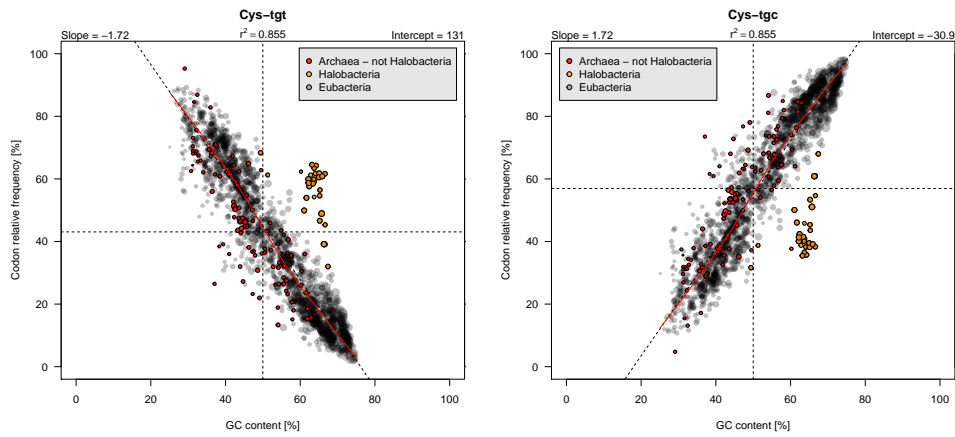
4.5.1 Asparagine



4.5.2 Aspartic acid

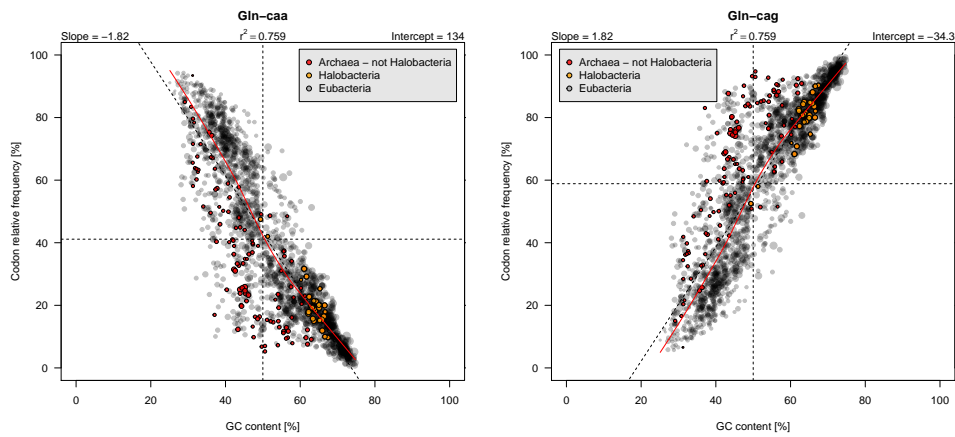


4.5.3 Cysteine

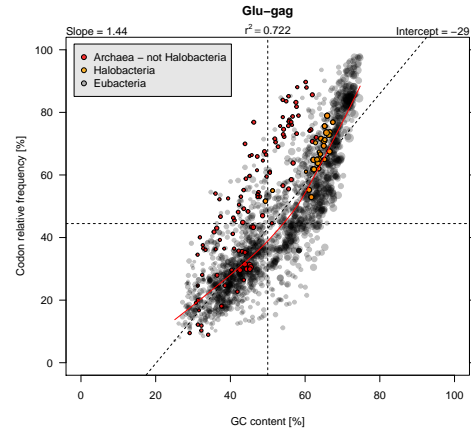
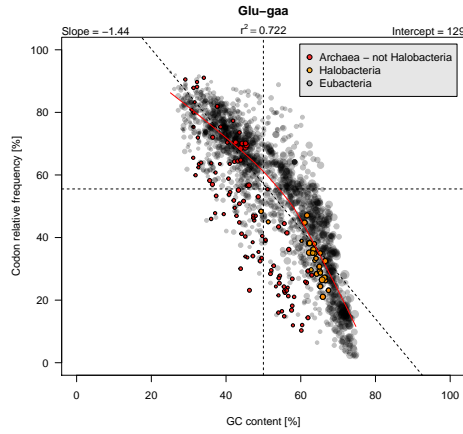


THE cysteine response is almost perfectly linear. The TGC codon is usually favored over the TGT codon. Halobacteria are an outlier group with a relative frequency of TGT much higher than expected from their high GC content.

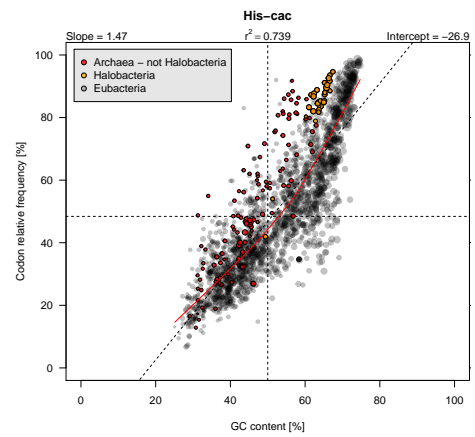
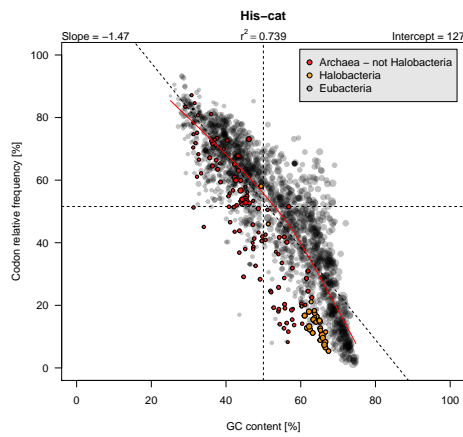
4.5.4 Glutamine



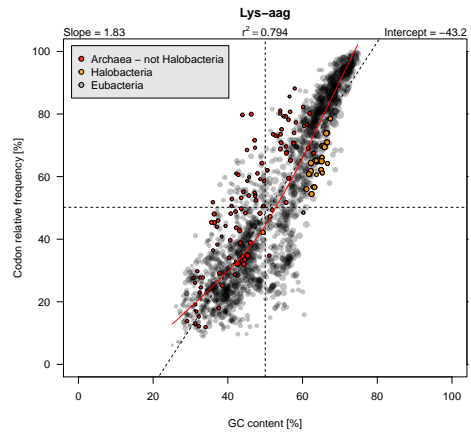
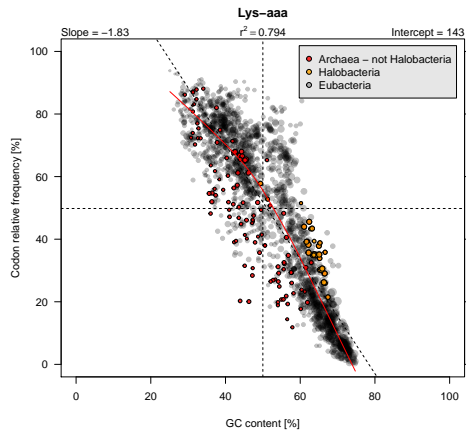
4.5.5 Glutamic acid



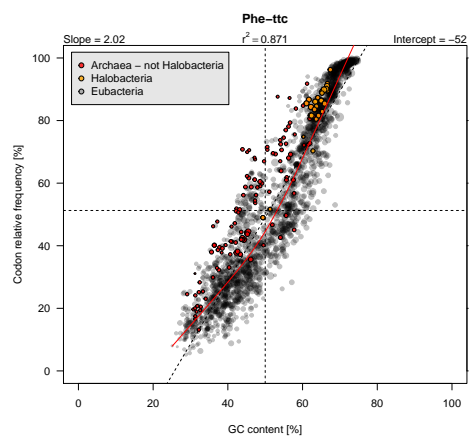
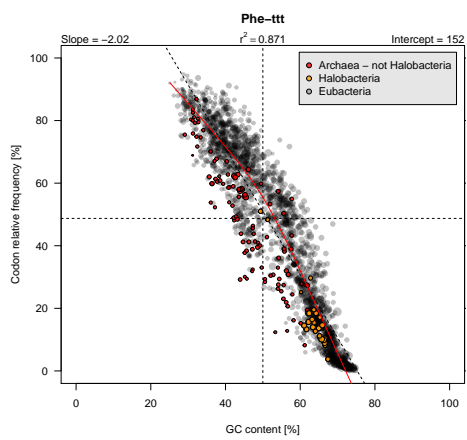
4.5.6 Histidine



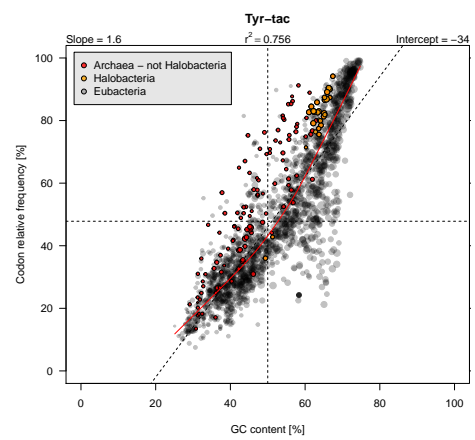
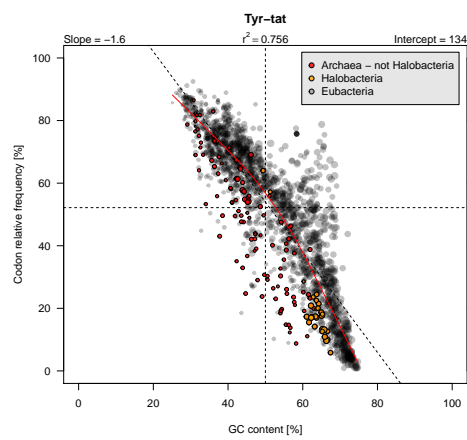
4.5.7 Lysine



4.5.8 Phenylalanine

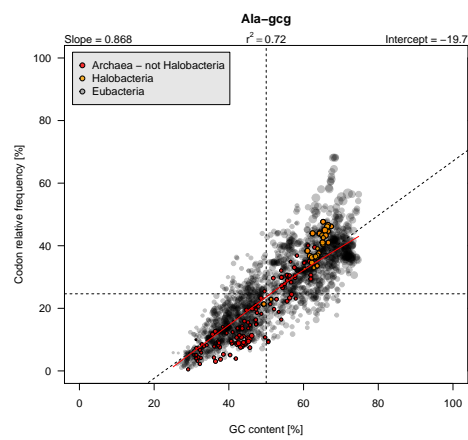
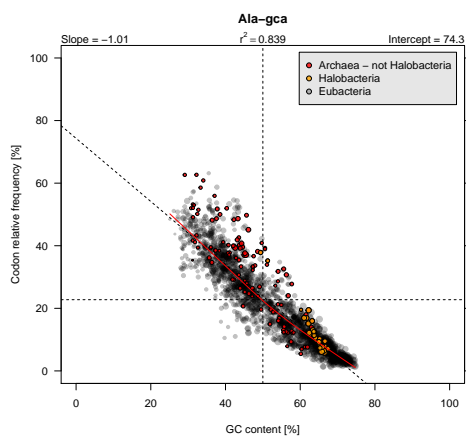
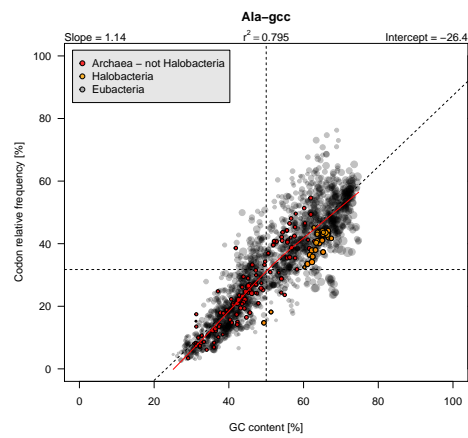
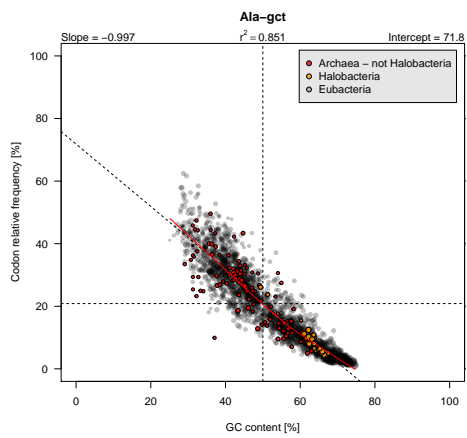


4.5.9 Tyrosine

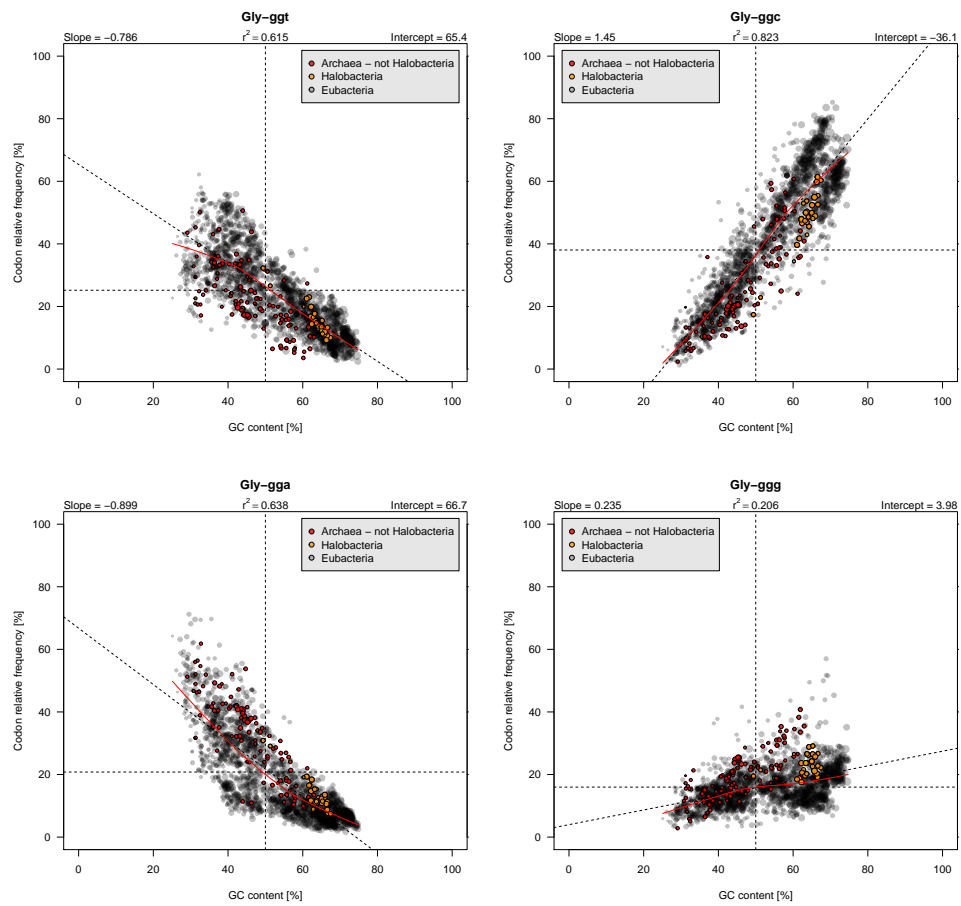


4.6 Quartet

4.6.1 Alanine

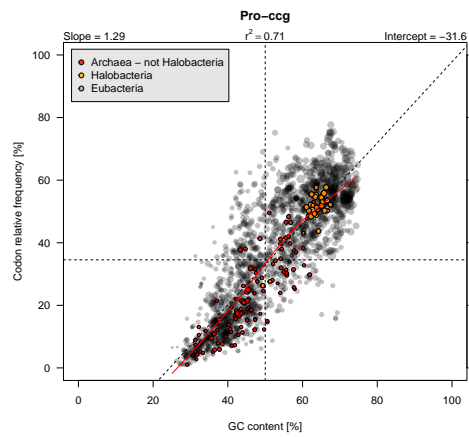
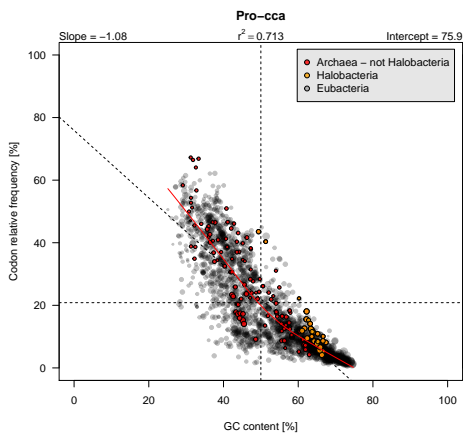
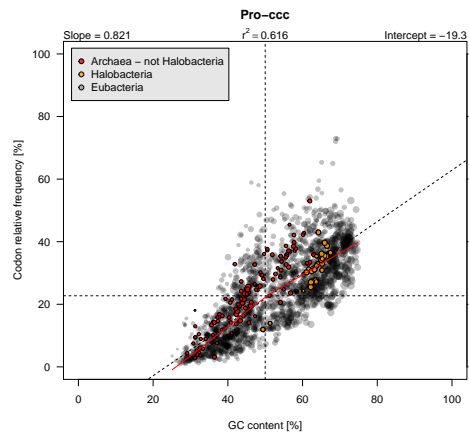
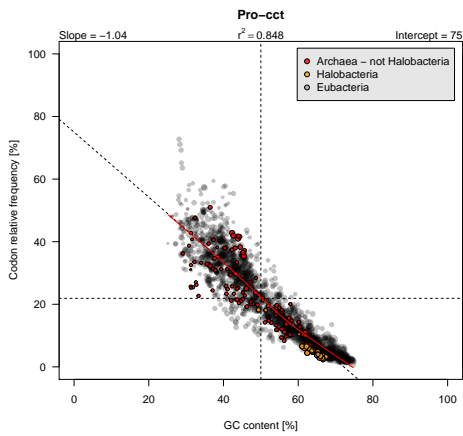


4.6.2 Glycine

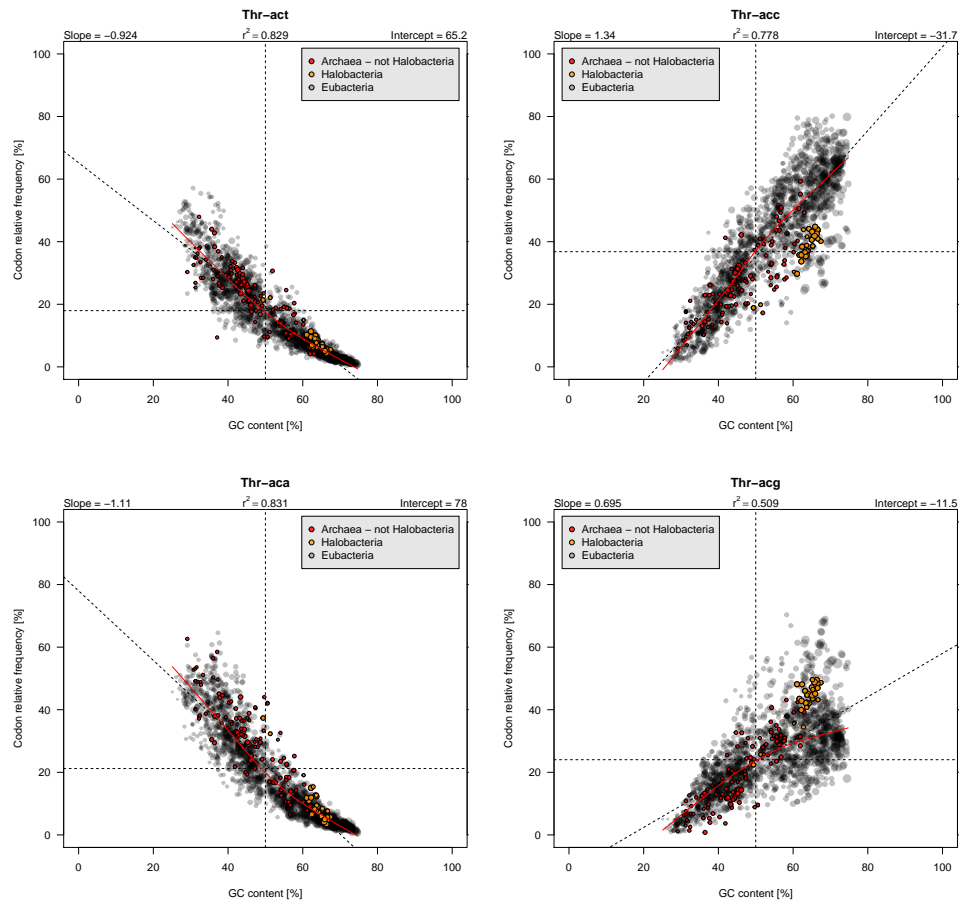


THE codon GGG seems to be counter-selected since even at very high GC its frequency is not so important.

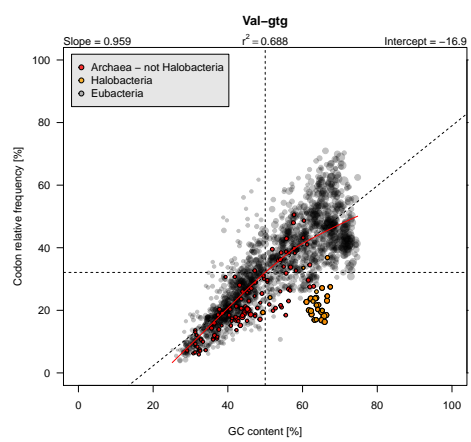
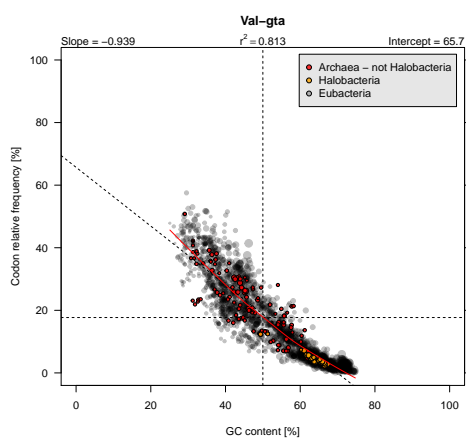
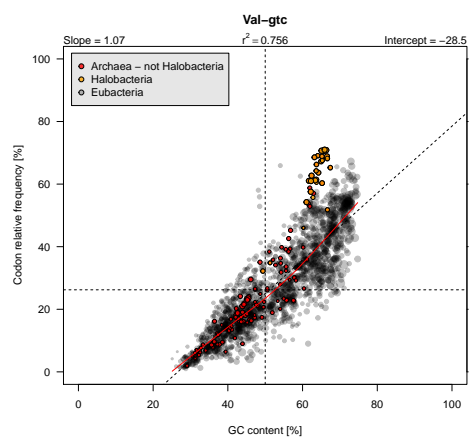
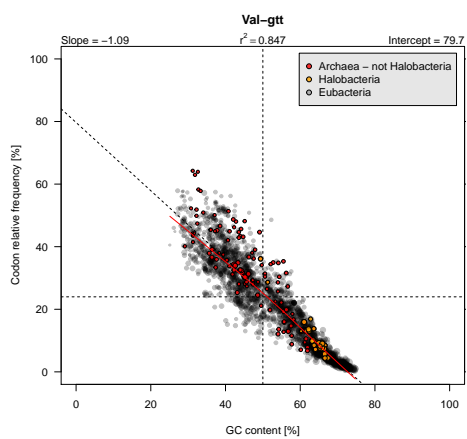
4.6.3 Proline



4.6.4 Threonine

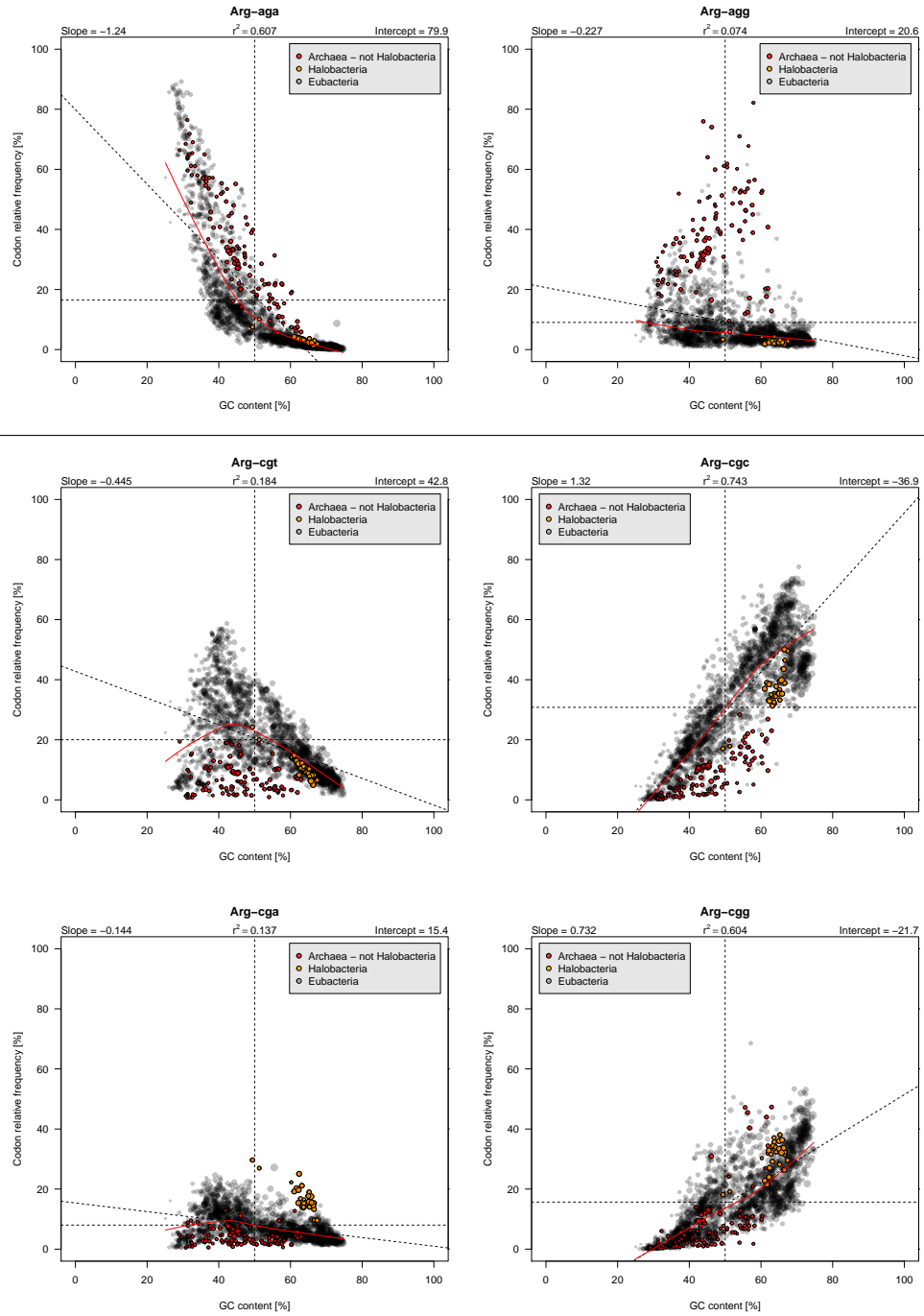


4.6.5 Valine



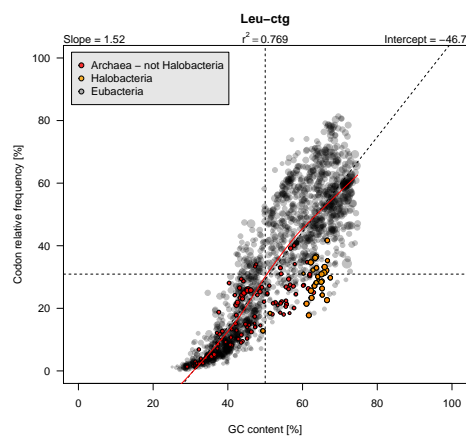
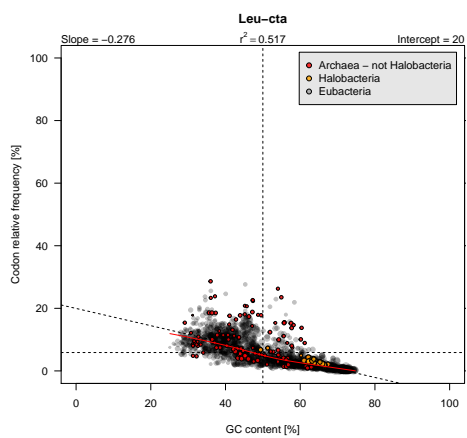
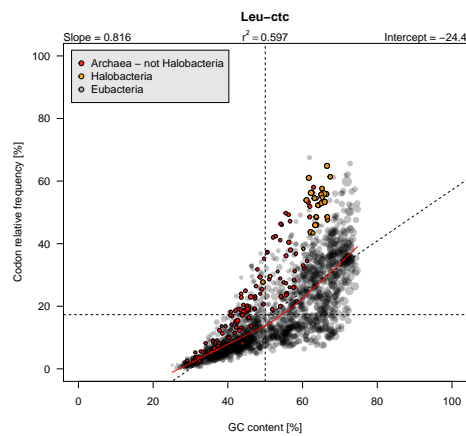
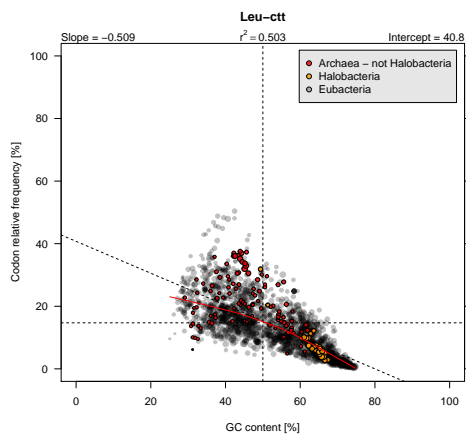
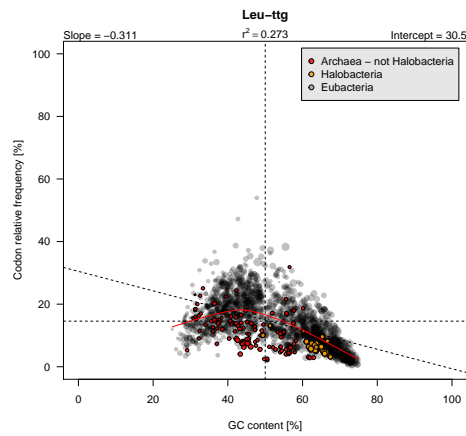
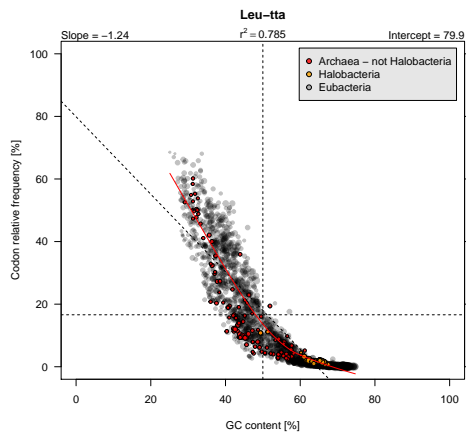
4.7 Sextet

4.7.1 Arginine

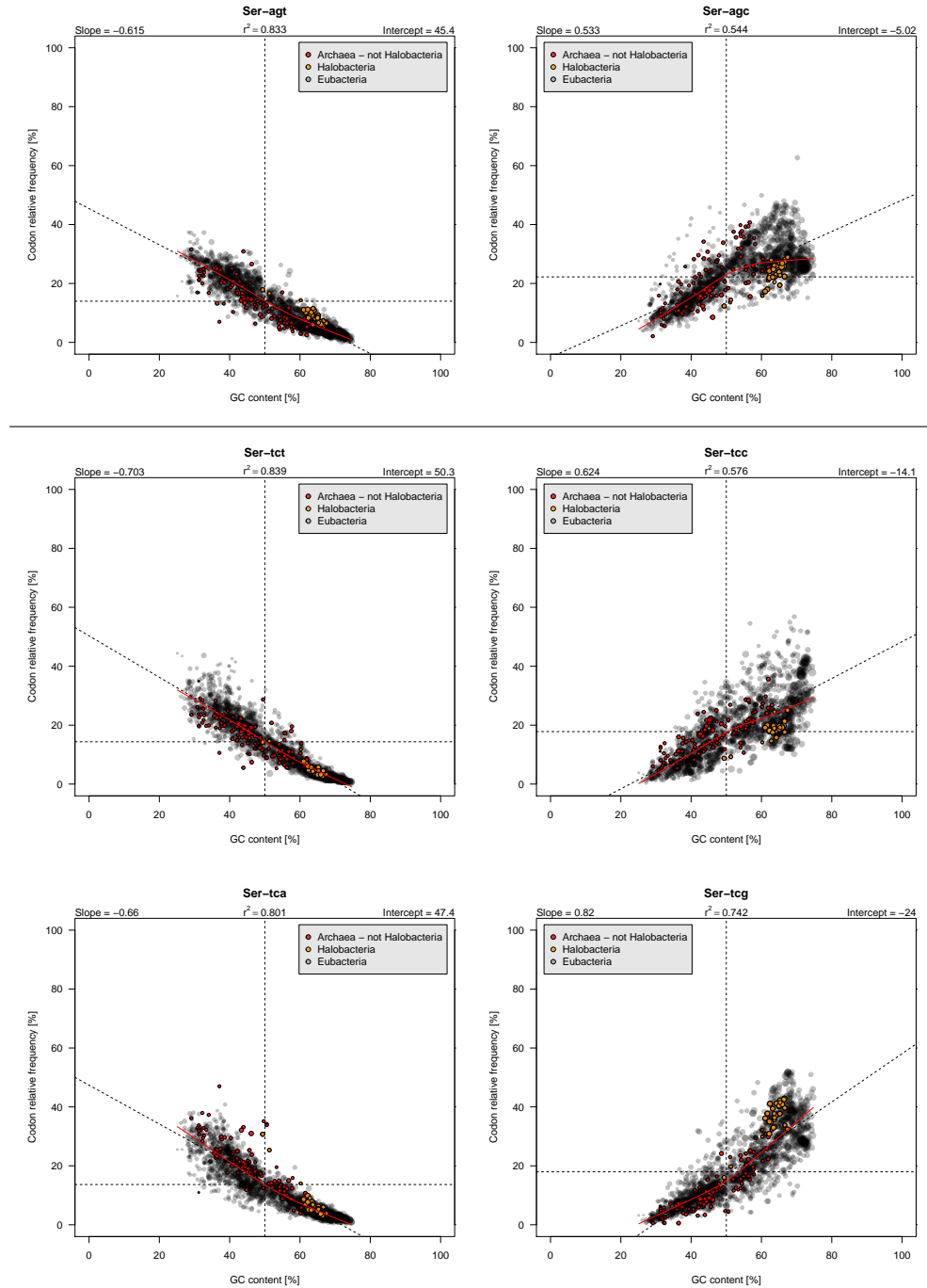


Rich pattern to comment here

4.7.2 Leucine



4.7.3 Serine



Chapter 5

Multivariate analysis of synonymous codon usage

5.1 Loading the dataset

```
load("local/tdd.Rda")
```


Chapter 6

Dataset compilation

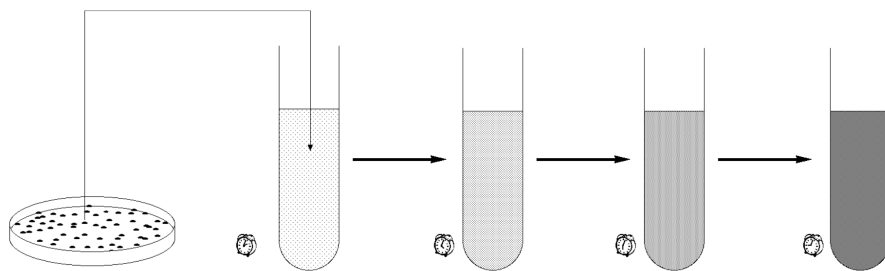
6.1 Introduction

6.1.1 Purpose

THIS chapter describes how the dataset used in the present book was obtained from various sources and then merged and curated in a single `data.frame` named `tdd`.

6.1.2 Bacterial growth as function of temperature

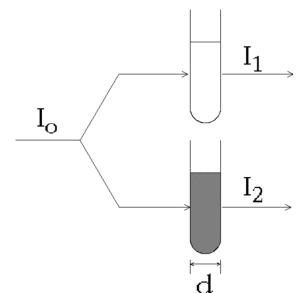
BECAUSE we are curating data for T_{opt} it's perhaps interesting to give some basic material here. You can skip this section and jump directly to section 6.2 page 101 if you are already familiar with this matter.



Is this section useful?
It is indeed convenient to introduce the `R` code for the CTMI model to avoid code duplication.

AS ILLUSTRATED just above in the case of a batch experiment a growth medium is inoculated with a colony isolated from a PETRI dish. Recording the darkness (aka turbidity) of the solution is a very popular technique to estimate biomass, that is the dry bacterial mass per volume unit (ML^{-3}). There is indeed a well established [84, 145, 122, 76, 130, 28, 58, 42] empiric law given in equation 6.1, analogous to the BEER-LAMBERT-BOUGUER law in chemistry, stating the proportionality between the bacterial contribution to the absorbance of the solution, A , and the biomass, B .

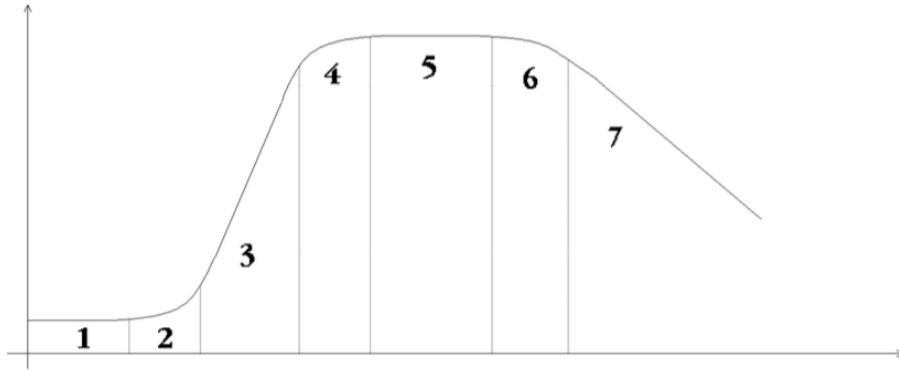
$$A = \log_{10} \frac{I_0}{I_2} - \log_{10} \frac{I_0}{I_1} = \log_{10} \frac{I_1}{I_2} = \alpha dB \quad (6.1)$$



AS DEPICTED in the margin, I_0 is the intensity of the incoming light, I_1 the intensity of the transmitted light without biomass (optical blank), I_2 the intensity of the transmitted light with biomass, d the length of the optical path and α a proportionality constant. The absorbance is often expressed for an optical path of 1 cm to define the optical density (OD) of the growth medium:

$$\text{OD} = \frac{1}{d}A = \alpha B \quad (6.2)$$

THE picture just below is the standard growth curve found in microbiology textbook established by BUCHANAN in 1918 [19]. This is a semi-logarithmic representation so the exponential growth phase labelled **3** is linear.



THE slope of the exponential growth phase in the semi-logarithmic representation is the specific growth rate, μ :

$$\mu = \frac{1}{B} \frac{dB}{dt} \quad (6.3)$$

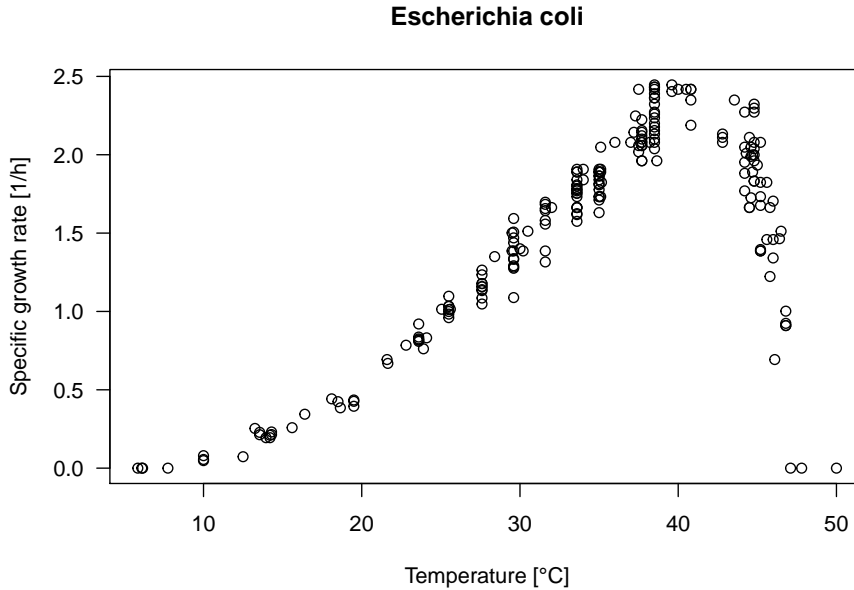
THE notion of cardinal temperatures was apparently first used in botany: I found the following in the April 1920 issue [99] of the *botanical gazette*:

CARDINAL TEMPERATURES.— As is well known, certain cardinal or fundamental temperatures are recognized. “Maximum” and “minimum” are terms used to refer to the highest and lowest temperatures at which the development of a particular organism may occur. The most favorable temperature for any process or function is designated the “optimum.”

THE oldest mention of *cardinal temperatures* I found is in a 1903 issue [98] of *The botanical gazette*, but again as a well known notion, without bibliographical references. The following figure illustrate the typical effect of temperature on μ with the data published in 1908 by BARBER [9] for *Escherichia coli*:

```
barber <- read.table(file="http://pbil.univ-lyon1.fr/R/donnees/barber.txt", header = TRUE)
save(barber, file = "local/barber.Rda")

load("local/barber.Rda")
plot(barber, las = 1, main = "Escherichia coli",
      xlab = "Temperature [°C]",
      ylab = "Specific growth rate [1/h]")
```



WE have then for the three cardinal temperatures $T_{\min} \approx 10^\circ\text{C}$, $T_{\text{opt}} \approx 40^\circ\text{C}$ and $T_{\max} \approx 50^\circ\text{C}$. Note that the curve is asymmetric, the following relationship holds:

$$T_{\text{opt}} > \frac{T_{\min} + T_{\max}}{2} \quad (6.4)$$

MORE accurate estimates are obtained by fitting a curve to the points. There are at least 10 published models for this purpose¹ from which the so-called square-root model from David RATKOWSKY [110] is the most popular (*cf.* figure 6.1 page 97). The CTMI (an acronym for Cardinal Temperature Model with Inflection) model [118] fit equally well the data [118, 148, 43, 44] but there is a structural correlation between the a and b parameter in the square-root while none is observed for the CTMI parameters. This allows easier estimation of the CTMI parameters whose biological interpretation is in addition straightforward. It didn't escape our notice that in a paper [111] titled *Empirical model with excellent statistical properties for describing temperature-dependent developmental rates of insects and mites* RATKOWSKY is now using the CTMI model²:

¹Recent (2017) review in [44]

²I was once (2017-03-02) asked by David RATKOWSKY: "It would be of interest to me if you could clarify for me the history behind the development of the LOBRY-ROSSO-FLANDROIS model. My colleagues and I have never been sure of who contributed what to the development of that model." Here was my (2017-04-08) answer: "here is what I remember from the history behind the development of the CTMI model. Take this *cum grano salis*. Here is a translation of a footnote page 85 from my PhD thesis [69] about the CTM (not the CTMI) model: 'This purely descriptive model was initially developed to illustrate the importance of the effect of choosing a model on parameter confidence limits (LOBRY *et al.* 1991 [74]). This model is far from being perfect because it can't take into account the inflection point which is often observed at low temperatures.' Modelling the effect of temperature on bacterial growth wasn't a core subject of my PhD. My concerns were from a methodological point of view as stated in this footnote to show that goodness of fit is not the unique important criterium and from a practical point of view to choose a realistic error model because in my experiments the

$$\begin{cases} \mu(T) = 0 & \text{if } T < T_{\min} \\ \mu(T) = \mu_{\text{opt}}\phi(T) & \text{if } T_{\min} \leq T \leq T_{\max} \\ \mu(T) = 0 & \text{if } T > T_{\max} \end{cases} \quad (6.5)$$

with

$$\phi(T) = \frac{\text{Num}(T)}{\text{Den}(T)} = \frac{(T - T_{\max})(T - T_{\min})^2}{(T_{\text{opt}} - T_{\min})[(T_{\text{opt}} - T_{\min})(T - T_{\text{opt}}) - (T_{\text{opt}} - T_{\max})(T_{\text{opt}} + T_{\min} - 2T)]} \quad (6.6)$$

THE corresponding `R` code is given below. Temperature, T , is the argument `Te` of the function `CTMI()` and `param` is a vector for the 4 parameters with `param[1]` for T_{\min} , `param[2]` for T_{opt} , `param[3]` for T_{\max} and `param[4]` for μ_{opt} . The variables `Num` and `Den` correspond to $\text{Num}(T)$ and $\text{Den}(T)$ in equation 6.6, respectively.

```
CTMI <- function(Te, param){
  Tmin <- param[1] ; Topt <- param[2] ; Tmax <- param[3] ; Muopt <- param[4]
  if( Te < Tmin || Te > Tmax ) return(0)
  Num <- (Te - Tmax)*(Te - Tmin)^2
  Den <- (Topt - Tmin)*((Topt - Tmin)*(Te - Topt) - (Topt - Tmax)*(Topt + Tmin - 2*Te))
  return(Muopt*Num/Den)
}
```

WITH the following `R` code we can now illustrate easily what the three cardinal temperatures are:

```
x <- seq(from = 0, to = 60, length.out = 500)
y <- sapply(x, CTMI, param = c(10, 40, 50, 2.5))
plot(x, y, type = "l", col = "darkblue", lwd = 2, las = 1,
      xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
      main = "The three cardinal temperatures")
arrows(10, 1, 10, 0.05, lwd = 2, angle = 15, length = 0.1)
text(10, 1, expression(italic(T)[min]), pos = 3, cex = 2)
arrows(40, 1.5, 40, 2.45, lwd = 2, angle = 15, length = 0.1)
text(40, 1.5, expression(italic(T)[opt]), pos = 1, cex = 2)
arrows(55, 1, 50.5, 0.05, lwd = 2, angle = 15, length = 0.1)
text(55, 1, expression(italic(T)[max]), pos = 3, cex = 2)
```

temperature was poorly controlled ($35.5 \pm 0.5^\circ\text{C}$ see figure on the top of page 87). I remember this as a quick-and-dirty modelling : I was familiar with rational functions $P(x)/Q(x)$ so that my approach was something like: Ok let's put a second degree for $P(x)$ to have a parabolic-shaped response and a first degree for $Q(x)$ so as to put a vertical asymptote on the right to make this asymmetric. Then it's just a trivial matter of parameter redefinition to introduce the cardinal temperatures and the maximum growth temperature. The CTM model was enough for my PhD but Laurent's concerns were such that temperature effect on bacterial growth was a core subject of his PhD. I can't tell you who did what exactly because it was a very close collaboration. I remember that we collected a lot of data from figures giving growth rate versus temperature. I remember that we increased the degree of $P(x)$ so as to have a cubic on top that can take into account the inflexion point, yielding the CTMI model (Cardinal Temperature Model with Inflexion point) published in JTB [118]."

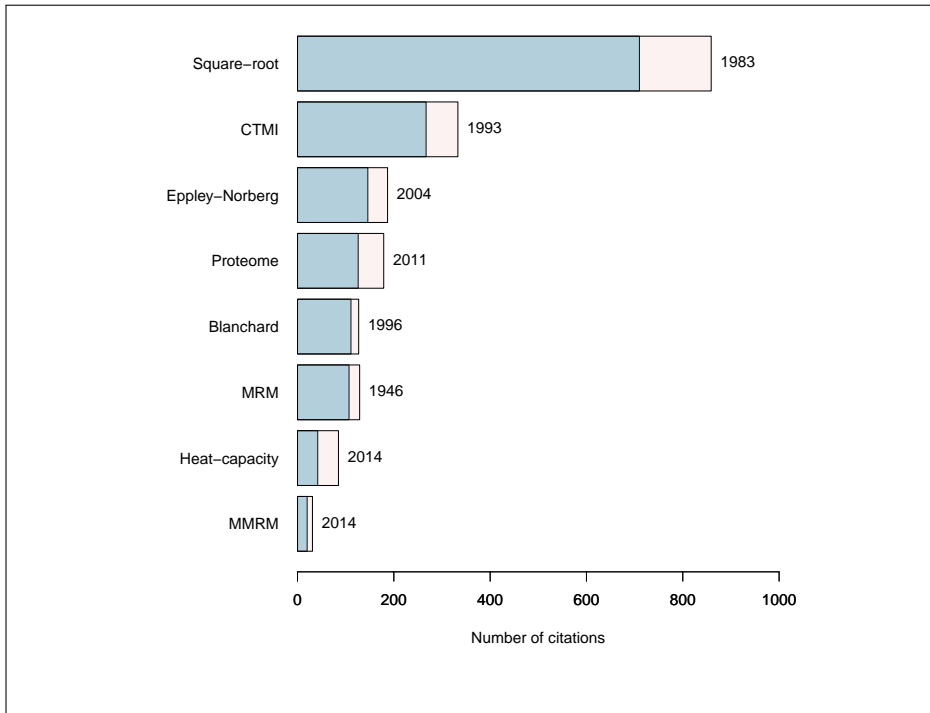

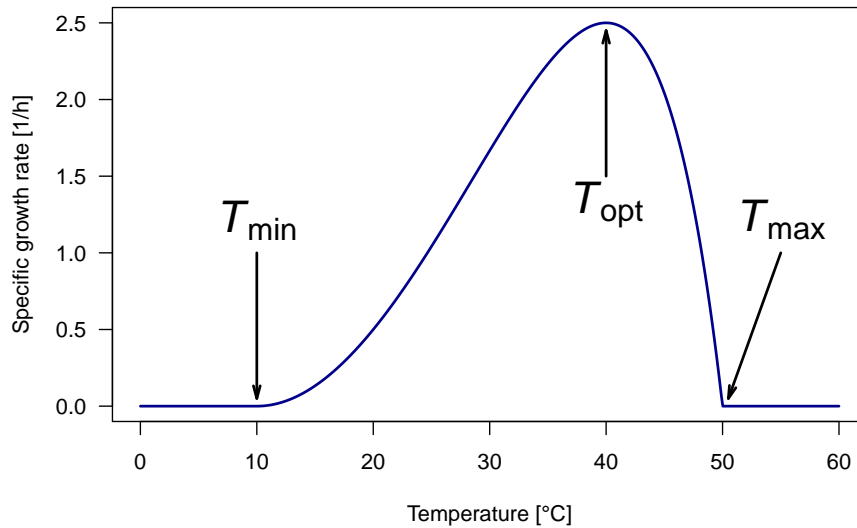


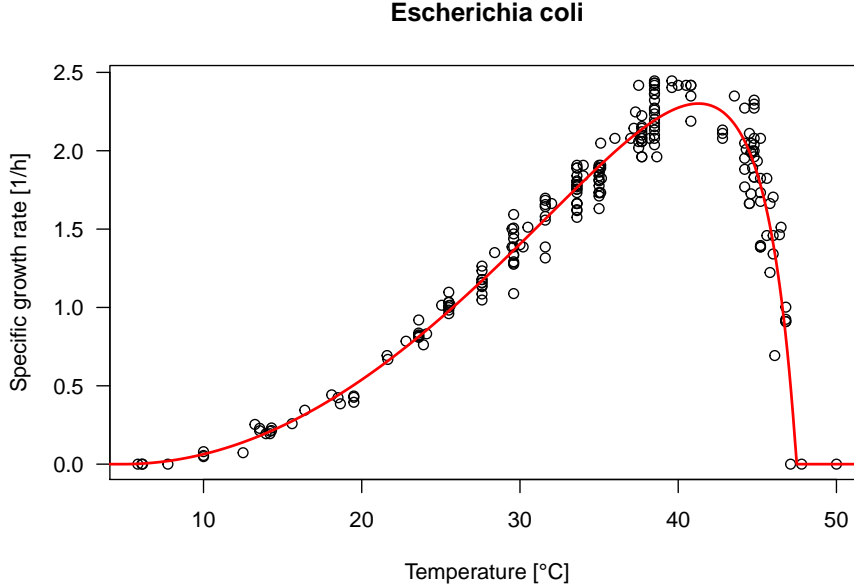
Figure 6.1: Popularity for 8 models predicting the specific growth rate, μ , as function of temperature. Data in blue are from table 2 in the recent (2017) review [44] and the pink extension is an update on 2019-05-24. The number of citations is estimated using “Google Scholar” citations for the first article presenting the model: Square-root *aka* RATKOWSKY [110], CTMI [118], EPPLEY-NORBERG [93], Proteome [29], BLANCHARD [15], MRM [52], Heat-capacity [123], MMRM [24]. The year of publication for the first article defining the model is given on the right of the bars. The DEB model [60] is not documented in [44] most likely because this is a general book that may be cited for many reasons others than the model itself. I have deleted the entry for the HINSHELWOOD model [49] because 507 citations are given in [44] but I found only 13 in my update. The  code for this figure is given p. 147.

The three cardinal temperatures



To fit the model to data we define the sum of squared residuals and then use the standard `nlm()` [R](#) built-in function to minimize its value. The vector `p` in `nlm()` arguments is the initial guess for parameter values which is set here directly from visual inspection of data.

```
sceCTMI <- function(param, data){
  xobs <- data[, 1]
  yobs <- data[, 2]
  ytheo <- sapply(xobs, CTMI, param)
  return( sum((yobs - ytheo)^2) )
}
nlm.barber <- nlm(sceCTMI, p = c(10, 40, 50, 2.5), data = barber)
load("local/barber.Rda")
plot(barber, las = 1, main = "Escherichia coli",
      xlab = "Temperature [°C]",
      ylab = "Specific growth rate [1/h]")
x <- seq(from = 0, to = 60, length.out = 500)
y <- sapply(x, CTMI, param = nlm.barber$estimate)
points(x, y, type = "l", col = "red", lwd = 2)
```



THE parameter estimates for *E. coli* with the BARBER dataset [9] are then with 3 significant digits: $T_{\min} = 4.89^{\circ}\text{C}$, $T_{\text{opt}} = 41.3^{\circ}\text{C}$, $T_{\max} = 47.5^{\circ}\text{C}$ and $\mu_{\text{opt}} = 2.3 \text{ h}^{-1}$. We could use this approach to estimate T_{opt} but it can be greatly improved by using the reparametrisation given by equation 7 in BERNARD and RÉMOND [12] to enforce the asymmetry 6.4.

$$T_{\max}(\eta) = T_{\text{opt}} + \frac{\eta^2}{\eta^2 + T_{\text{opt}}^2} (T_{\text{opt}} - T_{\min}) \quad (6.7)$$

I may miss something obvious here, but I don't understand the interest of the T_{opt}^2 factor (scaling factor?) in 6.7, let's simplify this to:

$$T_{\max}(\eta) = T_{\text{opt}} + \frac{\eta^2}{\eta^2 + 1} (T_{\text{opt}} - T_{\min}) \quad (6.8)$$

For both 6.7 and 6.8 we have:

$$\lim_{\eta \rightarrow \pm\infty} T_{\max}(\eta) = 2T_{\text{opt}} - T_{\min} \quad (6.9)$$

which ensures for any finite η that $T_{\max}(\eta) < 2T_{\text{opt}} - T_{\min}$, that is $T_{\max} - T_{\text{opt}} < T_{\text{opt}} - T_{\min}$, the required asymmetric feature 6.4. When $\eta = 0$ we have:

$$T_{\max}(0) = T_{\text{opt}} \quad (6.10)$$

In this case $\phi(T)$ in equation 6.6 degenerates to a simple parabolic function:

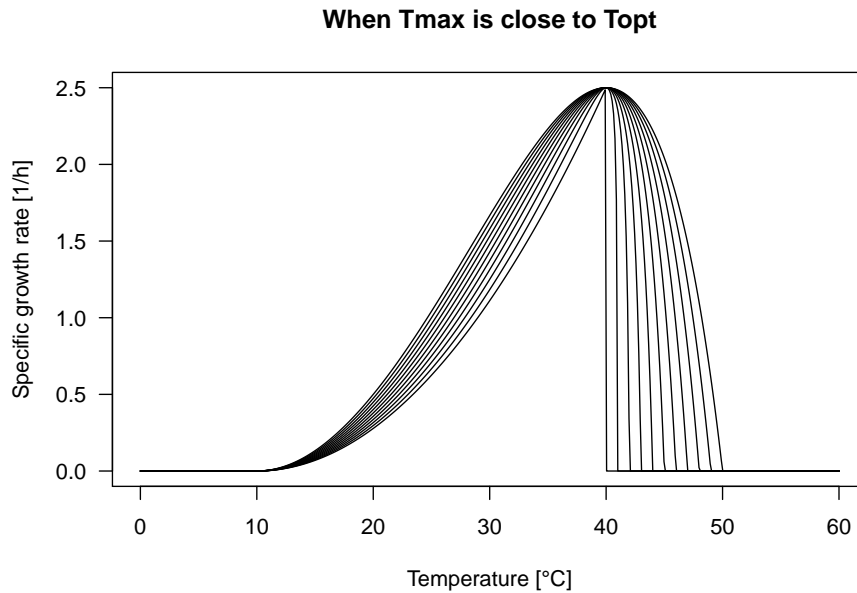
$$\phi(T|T_{\max} = T_{\text{opt}}) = \frac{(T - T_{\min})^2}{(T_{\max} - T_{\min})^2} \quad (6.11)$$

which injected into 6.5 turns it into a pure square-root model between T_{\min} and T_{\max} , ending at μ_{opt} :

$$\phi(T_{\max}|T_{\max} = T_{\text{opt}}) = \frac{(T_{\max} - T_{\min})^2}{(T_{\max} - T_{\min})^2} = 1 \quad (6.12)$$

Here is a graphical illustration of the degenerate case:

```
x <- seq(from = 0, to = 60, length.out = 500)
y <- sapply(x, CTMI, param = c(10, 40, 50, 2.5))
plot(x, y, type = "l", lwd = 1, las = 1,
      xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
      main = "When Tmax is close to Topt")
for(Tmax in seq(from = 49, to = 40, by = -1)){
  y <- sapply(x, CTMI, param = c(10, 40, Tmax, 2.5))
  points(x, y, type = "l")
}
```



From 6.8 we deduce:

$$\eta = \sqrt{\frac{T_{\max} - T_{\text{opt}}}{2T_{\text{opt}} - T_{\min} - T_{\max}}} \quad (6.13)$$

Now the **R** code to run this. The idea is to derive the function `nlm.CTMI()` from `nlm()` so that thanks to the dot-dot-dot argument all options available from `nlm()` are automatically handled by `nlm.CTMI()` too. The reparametrization is used internally so that the user works with the three cardinal temperatures and μ_{opt} to provide initial parameter guesses.

```
nlm.CTMI <- function(p , data, ...){
  CTMI <- function(Te, param){
    Tmin <- param[1] ; Topt <- param[2] ; eta <- param[3] ; Muopt <- param[4]
    Tmax <- Topt + (Topt - Tmin)*eta^2/(1 + eta^2)
    if( Te < Tmin || Te > Tmax ) return(0)
    Num <- (Te - Tmax)*(Te - Tmin)^2
    Den <- (Topt - Tmin)*((Topt - Tmin)*(Te - Topt) - (Topt - Tmax)*(Topt + Tmin - 2*Te))
    return(Muopt*Num/Den)
  }
  scele <- function(param, data){
```

```

  xobs <- data[ , 1] ; yobs <- data[ , 2]
  ytheo <- sapply(xobs, CTMI, param)
  return( sum((yobs - ytheo)^2) )
}
eta <- sqrt((p[3] - p[2])/(2*p[2] - p[1] - p[3]))
p[3] <- eta
res <- nlm(sceCTMI, p = p, data = data, ...)
res$internalestimate <- res$estimate
x <- res$estimate
Tmax <- x[2] + (x[2] - x[1])*x[3]^2/(1 + x[3]^2)
res$estimate[3] <- Tmax
return(res)
}
nlm.barber2 <- nlm.CTMI(p = c(10, 40, 50, 2.5), data = barber)
nlm.barber2$estimate
[1] 4.888177 41.282438 47.483222 2.301047
nlm.barber2$estimate
[1] 4.888053 41.282491 47.483188 2.301050
all.equal(nlm.barber2$estimate, nlm.barber2$estimate)
[1] "Mean relative difference: 2.241853e-06"


```

AN OTHER possible improvement is to automatically set the initial guess from data inspection. Here T_{\min} is set from the minimal temperature reported in the dataset, T_{\max} from maximal temperature reported in the dataset, μ_{opt} from the maximal specific growth rate in the dataset and T_{opt} to its corresponding temperature.

```


nlm.CTMI.auto <- function(data, ...){
  p <- numeric(4)
  p[1] <- min(data[ , 1]) ; p[3] <- max(data[ , 1])
  i <- which.max(data[ , 2])
  p[2] <- data[i, 1] ; p[4] <- data[i, 2]
  nlm.CTMI(p = p, data, ...)
}
nlm.barber3 <- nlm.CTMI.auto(data = barber)
all.equal(nlm.barber3$estimate, nlm.barber3$estimate)
[1] "Mean relative difference: 1.445609e-06"

```

NOTE that these initial estimates could be too crude for ill-conditioned data and make convergence impossible. In this case it's better to turn back to the `nlm.CTMI()` function to better control initial guesses. Some examples of this situation are visible in the  code for figure 6.11 page 120, figure 6.8 page 117 and figure 6.3 page 109.

6.2 Origin of data

6.2.1 T_{opt} data from Engqvist 2018

THIS dataset was mined [31] by Martin K. M. ENGQVIST in 2018 from the web sites of 5 microbial culture collections, *viz.*³, along with BacDive [140]. With 21,498 documented species this is to date the most complete source of optimal growth temperatures. Data⁴ were directly imported under  as follows. They were saved then in XDR [139] format to allow for off-line work.

³ATCC for the United States of America, DSMZ for the Federal Republic of Germany, NCTC for the United Kingdom of Great Britain and Northern Ireland, NIES for Japan, CIP for the French Republic.

⁴<https://doi.org/10.5281/zenodo.1175608>

Say something about the strange correlation between the 3 cardinal temperatures?

Say something about the non consensual classification into psy-meso-thermo-hyperthermophiles?

```

path <- "https://zenodo.org/record/1175609/files/temperature_data.tsv"
MKME <- read.table(path, sep = "\t", header = TRUE, stringsAsFactors = FALSE)
save(MKME, file = "local/MKME.Rda")

load("local/MKME.Rda")
dim(MKME)
[1] 21498  11
names(MKME)
[1] "organism"    "domain"      "temperature" "taxid"        "lineage_text"
[6] "superkingdom" "phylum"    "class"        "order"        "family"
[11] "genus"

```

Describe variables in MKME

6.2.2 Codon usage data from Lobry 2018

Re-run with internet on

```

path <- "http://pbil.univ-lyon1.fr/R/donnees/JLO/stabilty.rda"
load(url(path))
save(bact, file = "local/bact.Rda") # codon usage table for 12,317 strains
save(topt, file = "local/topt.Rda") # Topt for 740 strains (LN2006)

```

```
load("local/bact.Rda")
```

THE dataset used for figure 1.10 in [71] was retrieved from release 7 of hogenom [102]. It consists of 13,165,776,353 codon counts from 12,317 bacterial strains representing 2,301 species and 980 genera. It is therefore a contingency table in which codons are distributed among the crossed levels between the 107 codons and the 1883 bacterial strains. An important limitation of this dataset is that all bacterial using a non-standard genetic code have been removed. This is mandatory to study synonymous codon usage but not necessary for aminoacid usage study.

Give the list of species with non-standard genetic code

THERE is a problem in this dataset because of the single quote in the ASTER YELLOWS WITCHES'-BROOM PHYTOPLASMA entry that messed the import. Here is a fix starting back to the text file:

```

# Show the problem, one string is too long for a species name:
tail(sort(nchar(bact$species)))
[1] 79 80 80 80 85 10894
# Import command was:
# bact <- read.table("bacteria.out", sep = "\t", stringsAsFactors = FALSE)
# The fix is:
bact <- read.table("local/bacteria.out", sep = "\t", stringsAsFactors = FALSE,
quote = '')
colnames(bact) <- c("species", "nCDS", "group", words())
tail(sort(nchar(bact$species))) # OK now
[1] 77 79 80 80 80 85
bact$nCDS <- as.numeric(bact$nCDS)
for(j in 4:67) bact[, j] <- as.numeric(bact[, j])
# Show deleted entries
bact[bact$nCDS < 250, "species"]
[1] "CANDIDATUS PARVARCHAEUM ACIDIPHILUM ARMAN-4_'5-WAY FS'"
[2] "CANDIDATUS PARVARCHAEUM ACIDOPHILUS ARMAN-5_'5-WAY FS'"
[3] "PREVOTELLA ORYZAE DSM 17970"
[4] "CANDIDATUS SULCIA MUELLERI GWSS"
[5] "CANDIDATUS SULCIA MUELLERI SMDSEM"
[6] "CANDIDATUS SULCIA MUELLERI DMIN"
[7] "CANDIDATUS SULCIA MUELLERI CARI"
[8] "CANDIDATUS SULCIA MUELLERI STR. SULCIA-ALF"

```

```
[9] "CANDIDATUS WALCZUCHELLA MONOPHLEBIDARUM"
[10] "CANDIDATUS UZINURA DIASPIDICOLA STR. ASNER"
[11] "CITROBACTER SP. S-77"
[12] "PHOTOBACTERIUM DAMSELAE SUBSP. PISCICIDA DI21"
[13] "PSEUDOMONAS AERUGINOSA PAK"
[14] "CANDIDATUS CARSONELLA RUDDII PV"
[15] "CANDIDATUS CARSONELLA RUDDII CE ISOLATE THA02000"
[16] "CANDIDATUS CARSONELLA RUDDII CS ISOLATE THA02000"
[17] "CANDIDATUS CARSONELLA RUDDII HC ISOLATE THA02000"
[18] "CANDIDATUS CARSONELLA RUDDII HT ISOLATE THA02000"
[19] "CANDIDATUS CARSONELLA RUDDII PC ISOLATE NHV"
[20] "CANDIDATUS CARSONELLA RUDDII DC"
[21] "CANDIDATUS ZINDERIA INSECTICOLA CARI"
[22] "CANDIDATUS TREMBLAYA PRINCEPS PCIT"
[23] "CANDIDATUS TREMBLAYA PRINCEPS PCVAL"
[24] "CANDIDATUS TREMBLAYA PHENACOLA PAVE"
[25] "CANDIDATUS NASUTIA DELTOCEPHALINICOLA"
[26] "SPHINGOMONAS JASPSI DSM 18422"
[27] "DESULFOVIBRIO TERMITIDIS H11"
[28] "ACTINOSPICA ROBINIAE DSM 44927"
[29] "LEPTOLYNGBYA VALDERIANA BDU 20041"

# Delete entries
bact <- bact[bact$nCDS >= 250, ] # at least 250 CDS
save(bact, file = "local/bact2.Rda") # codon usage table for 12,345 strains
# That is 28 more than in the previous version
```

THE following `R` code is to compute species names with the same format as in the column `organism` in data from [31], that is for instance `abiotrophia_adiacens`. Do not end with a dot for *Genus* sp.

```
csspn <- function(x){
  tmp <- tolower(x)
  tmp <- unlist(strsplit(tmp, split = " "))
  # remove [ ] and '
  for(i in 1:2){
    target <- c("[", " ", ".", ",")
    if(substr(tmp[i], 1, 1) %in% target){
      tmp[i] <- substr(tmp[i], 2, nchar(tmp[i]))
    }
    if(substr(tmp[i], nchar(tmp[i]), nchar(tmp[i])) %in% target){
      tmp[i] <- substr(tmp[i], 1, nchar(tmp[i]) - 1)
    }
  }
  return(paste(tmp[1], tmp[2], sep = "_"))
}
# Check with some cases:
csspn("CANDIDATUS KORIBACTER VERSATILIS ELLIN345")
[1] "candidatus_koribacter"
csspn("[BREVI BACTERIUM] FLAVUM")
[1] "brevibacterium_flavum"
csspn("[PSEUDOMONAS SYRINGAE] PV. TOMATO STR. DC3000")
[1] "pseudomonas_syringae"
csspn("'NOSTOC AZOLLAE' 0708")
[1] "nostoc_azollae"
csspn("HYDROGENOBACULUM SP. Y04AAS1")
[1] "hydrogenobaculum_sp"
# Compute for whole dataset
bact$organism <- sapply(bact$species, csspn)
save(bact, file = "local/bact2.Rda")

load("local/bact2.Rda")
sum(bact[, 4:67])
[1] 13182093496
```

IN this dataset the taxonomic leaves are at the strain level, so that we want to aggregate data at the species level so as to be able to merge with T_{opt} data. There are different ways to do this, here the average codon usage for all the strains of a given species was computed.

Explain why it's better than a simple sum

```
# aggregate data by species:
tocu <- apply(bact[, 4:67], 2, function(x) tapply(x, bact$organism,
function(x) round(sum(x)/length(x))))
tocu <- as.data.frame(tocu)
dim(tocu) # 2293 X 64
[1] 2293 64
tocu$organism <- rownames(tocu)
sum(tocu[, 1:64]) # 2,478,128,298
[1] 2478128298
```

Fix me!

TO sum up a table of codon usage data for 2293 bacterial species is at hand here. The `R` code used here wouldn't be harmed by more explanations.

6.2.3 T_{opt} data from Lobry & Necsulea 2006

THIS dataset [73] contains T_{opt} data for 740 bacterial strains from 458 distinct species. These data are issued from [38] but reworked by comparing with the *Prokaryotic Growth Temperature database* [50]⁵ and to the DSMZ. They are then of a better quality but restricted to the species with genomic data available in 2006.

is the link still dead ?

```
load("local/topt.Rda")
topt$organism <- sapply(topt$species, csspn)
# aggregate data by species:
toptsp <- tapply(topt$toptmean, topt$organism, mean)
toptsp <- as.data.frame(toptsp)
toptsp$organism <- rownames(toptsp)
dim(toptsp) # 458 X 2
[1] 458 2
```

THE table `toptsp` has two columns: `toptsp` is the mean T_{opt} across all the strains of a given species, `organism` is the species as in [31]. There are 458 documented species here.

6.2.4 T_{opt} data from Galtier & Lobry 1997

WE want to add data from [38] that are not already present in [73]. The main source for [38] is a manual scan of *Bergey's manual* [133] (1984–1989) plus a compilation of hyperthermophiles archaea [25]. There are T_{opt} data for 772 bacterial strains from 766 distinct species.

```
GL <- read.table("ftp://pbil.univ-lyon1.fr/pub/datasets/JME97/species",
header = FALSE, sep = "\t")
save(GL, file = "local/GL.Rda")

load("local/GL.Rda")
GL$organism <- paste(GL$V1, GL$V2)
GL$organism <- sapply(GL$organism, csspn)
any(duplicated(GL$organism)) # TRUE
[1] TRUE
GL$organism[duplicated(GL$organism)] # 6 duplicated species but
[1] "beggiatoa_alba" "butyrivibrio_crossotus"
[3] "halobacterium_salinarium" "nocardioides_albus"
[5] "promicromonospora_citreus" "rhodococcus_marinonascens"
```

⁵The link <http://pgtdb.csie.ncu.edu.tw> given in the article is not responding (last consultation 2019-03-26) however [78] are still referring to it in 2016.


```
# with same Topt. Delete duplicates:
GL <- GL[!duplicated(GL$organism), ]
nrow(GL) # 766
[1] 766

new <- !(GL$organism %in% toptsp$organism)
sum(new) # 660
[1] 659

new.df <- as.data.frame(list(toptsp = GL$V4[new], organism = GL$organism[new]))
toptsp <- rbind(toptsp, new.df)
toptsp <- toptsp[order(toptsp$organism), ]
nrow(toptsp) # 1118
[1] 1117
```

WE have therefore T_{opt} data for 1117 bacterial species. We use a left join with the table of codon usage to merge data.

```
tocuT1 <- merge(tocu, toptsp, all.x = TRUE)
sum(!is.na(tocuT1$toptsp)) # 486
[1] 486
```

WE have a total of 2293 bacterial species with codon usage data and T_{opt} documented in [38, 73].

6.3 T_{opt} curation

6.3.1 Merging tables

A limitation [31] of data obtained from microbial culture collections is that given temperatures may not always correspond to genuine T_{opt} but simply to the temperature used to sustain the growth of organisms. To check the quality of data we want to compare with the temperatures from [38, 73] that are actual T_{opt} .

```
tocuT2 <- merge(tocuT1, MKME, all.x = TRUE)
```

6.3.2 Taxonomic filtering

I have removed all the *Candidatus* for two reasons:

1. *Candidatus* is used to prefix bacterial species, for instance “*Candidatus* Methanoplasma termitum”, that cannot be maintained in a microbiological culture collection because we don’t know how to grow them. There is no way to estimate T_{opt} in this case. For endosymbionts you may think using the host temperature as a proxy for T_{opt} but the psychrotrophic bacteria isolated from a constantly warm tropical environment [2] argues against this approach.
2. I have only the genus name in the data from [31], for instance *candidatus_amoebophilus*, so that I can’t make a clean join at the species level.

```
cand <- grep("candidatus", tocuT2$organism)
#tocuT2$organism[cand]
tocuT2 <- tocuT2[-cand, ]
```

I have also deleted the following entries that do not belong to the standard binominal nomenclature:

```
i_aster_yellows <- which(tocuT2$organism == "aster_yellows")
i_onion_yellows <- which(tocuT2$organism == "onion_yellows")
tocuT2 <- tocuT2[-c(i_aster_yellows, i_onion_yellows), ]
```

I didn't remove data documented only at the genus level, that is *Genus* sp., because that would delete more than 200 entries. Note that there are genus such as *Bacillus* [150] with a wide range of T_{opt} , from 25°C to 67°C, so that this perhaps not a good thing to do.

Idea: delete Genus sp. iff no Genus species available

```
issp <- function(x) substr(x, nchar(x) - 2, nchar(x)) == "_sp"
sum(sapply(tocuT2$organism, issp))
```

```
[1] 214
```

6.3.3 Available T_{opt} before curation

translate iff usefull. Mosaicplot?

```
tocuT2$genuine <- !(is.na(tocuT2$toptsp))
tocuT2$mined <- !(is.na(tocuT2$temperature))
n <- nrow(tocuT2)
(xor <- with(tocuT2, sum(genuine | mined)))
```

```
[1] 1874
```

```
(and <- with(tocuT2, sum(genuine & mined)))
```

```
[1] 441
```

```
(crossdoc <- with(tocuT2, table(genuine, mined)))
```

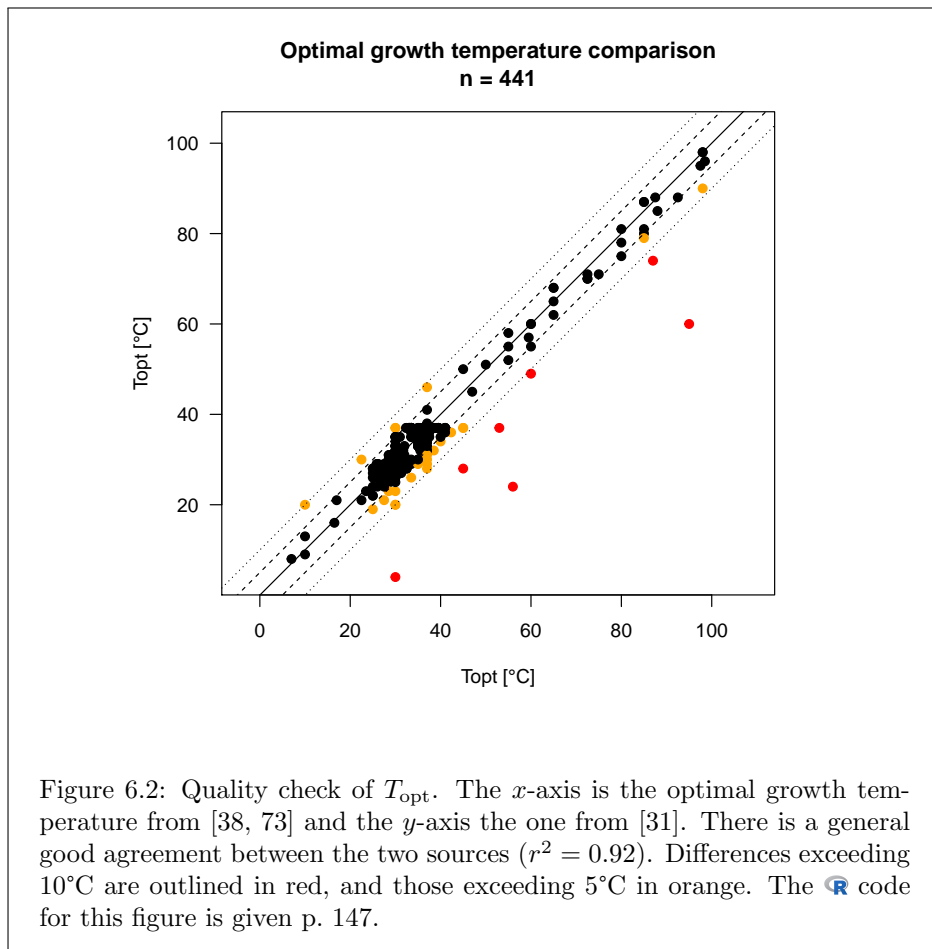
```
      mined
genuine FALSE TRUE
FALSE   358 1393
TRUE    40  441
```

```
#mosaicplot(crossdoc)
```

TO summarize, out of 2232 species I have at least one T_{opt} indication for 1874 species, that is 84%. I have 441 species with T_{opt} documented in both sources, allowing for comparisons.

6.3.4 T_{opt} comparison between [31] and [38, 73]

FIGURE 6.2 page 107 shows that T_{opt} are very consistent between [31] and [38, 73]. Out of 441 pairs of values, only 33 (7.48%) differ by more than 5°C and among those 7 (7.48%) differ by more than 10°C. In others words, 408 values (92.52%) differ by less than 5°C. The 441 manually curated T_{opt} values from [38, 73] represent admitly a small subset (2.05%) of the 21498 temperature data from [31], but it is reconforting to have such a good agreement here. We have now to solve the outliers.



6.3.5 Solving important T_{opt} discrepancies ($> 5^{\circ}\text{C}$)

IN this section I want to make some bibliographical searches and parameter estimations to solve important T_{opt} discrepancies. The list of species under study is given in table 6.1 page 111.

FOR *Thermococcus kodakarensis* it's a new species based on strain KOD1 from *Pyrococcus* sp. [3]. According to table 1 of this paper the range of temperature for growth is 60-100°C with an optimum at about 85°C. In [109] it is however stated that “the optimum temperature for enzyme activity was shown to be 60°C, although the optimum growth temperature of the strain KOD1 is 95°C” but without data nor references. In the abstract of the paper [86] describing the purification of strain KOD1 it is stated that “the growth temperature of the strain ranged from 65 to 100°C, and the optimal temperature was 95°C”. Inspection of Figure 1a shows that there is indeed an optimum at 95°C but the y -axis is a biomass density (in Cells/ml), not a specific growth rate (1/time). The 95°C is more likely an optimal temperature for growth yield, not for growth rate. I have selected $T_{\text{opt}} = 85^{\circ}\text{C}$ from [3].

```
tocuT2$topt <- NA # new column
tocuT2[which(tocuT2$organism == "thermococcus_kodakarensis"), "topt"] <- 85
```

FOR *Synechococcus elongatus* (NÄGELI) NÄGELI 1849 strain PCC6301 from table 1 in [77] T_{opt} is at least 35°C.

Complete *Synechococcus elongatus*

FOR *Shewanella violacea*, table 3 from [54] gives an optimal growth temperature equal to 8°C and refer to [92] for the description of strain DSS12 as a member of *Shewanella violacea* sp. nov. who refer to [55] for T_{opt} estimation. Data are in figure 2F for strain DSS12 with only two temperature conditions (8°C and 15°C) and an optimum at 30 MPa, and we know in addition from the text that it was not able to growth at temperature above 20°C. The optimal growth temperature is therefore not very accurately estimated, I have selected $T_{\text{opt}} = 8^{\circ}\text{C}$.

```
tocuT2[which(tocuT2$organism == "shewanella_violacea"), "topt"] <- 8
```

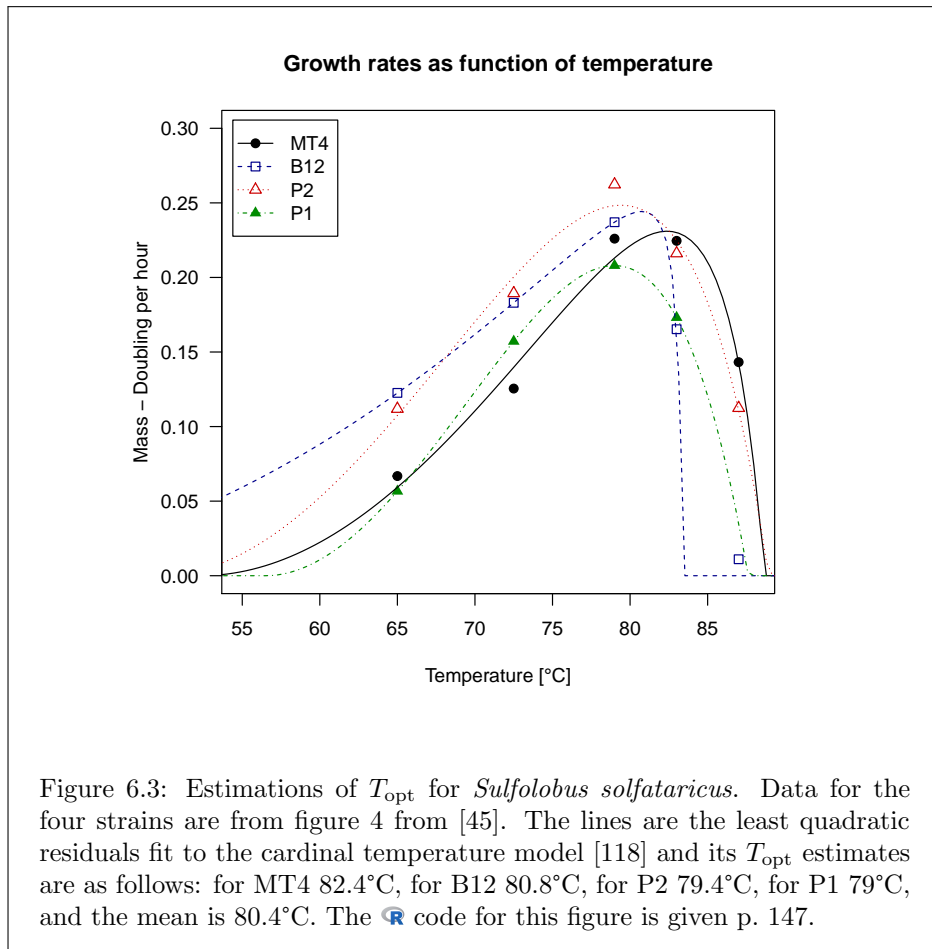
FOR *Streptomyces glaucescens* (strain DSM 40716) according to the introduction from [57] it's a mesophile with T_{opt} close to 25°C. Species description is given in [127] without T_{opt} mention. According to table 1 from [67], *Streptomyces glaucescens* (strain DSM 40155) is able to growth at 45°C. I wasn't able to find a paper with a documented experimental data for T_{opt} estimation, so I kept the 28°C from [31].

```
tocuT2[which(tocuT2$organism == "streptomyces_glaucescens"), "topt"] <- 28
```

FOR *Bacillus coagulans* (strain B666) according to table 2 from [150] T_{opt} equals 55°C, so that the 53°C from [38, 73] seems better than the 16°C from [31]. However, according to [89] *B. coagulans* has a wide range of T_{opt} , so I have deleted this entry.

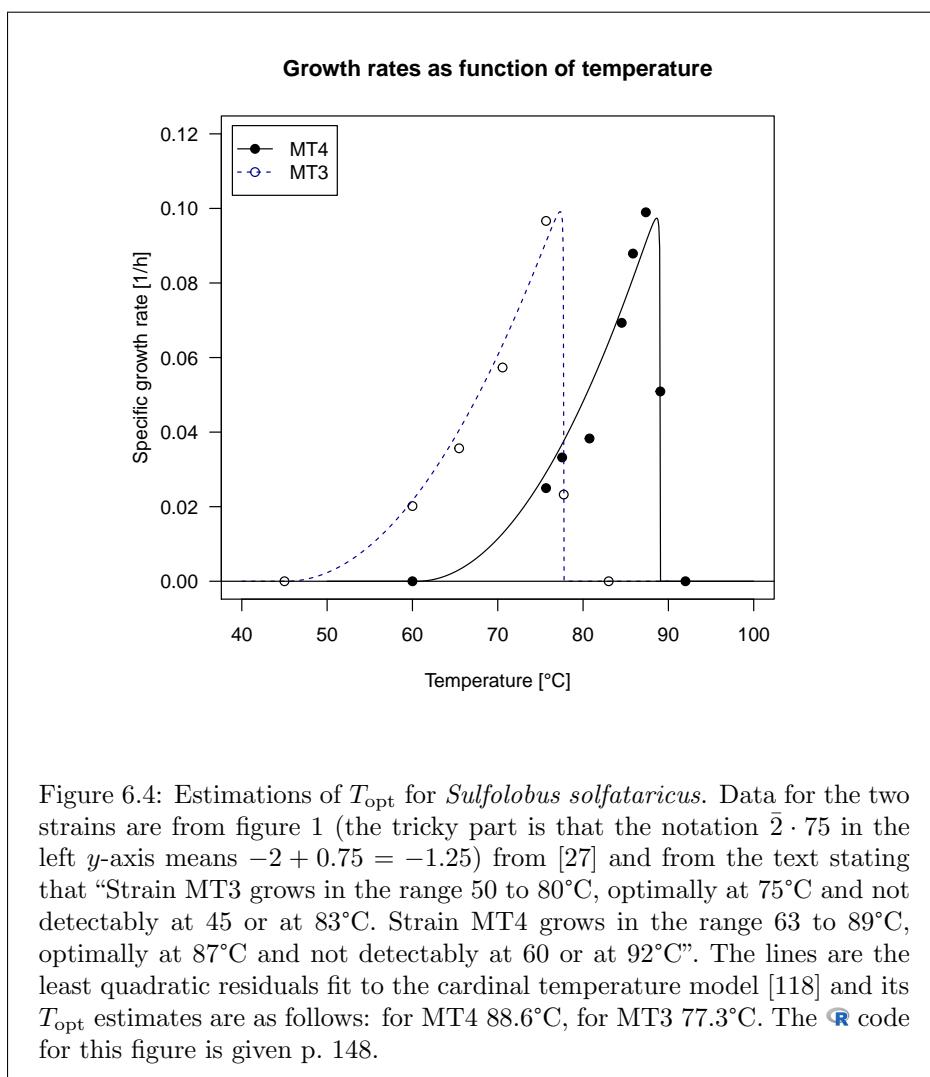
```
todelete <- which(tocuT2$organism == "bacillus_coagulans")
```

FOR *Sulfolobus solfataricus* (strains MT4, B12, P1, P2) according to figure 4 from [45] T_{opt} is in the range 80-85°C. This figure is re-created in figure 6.3



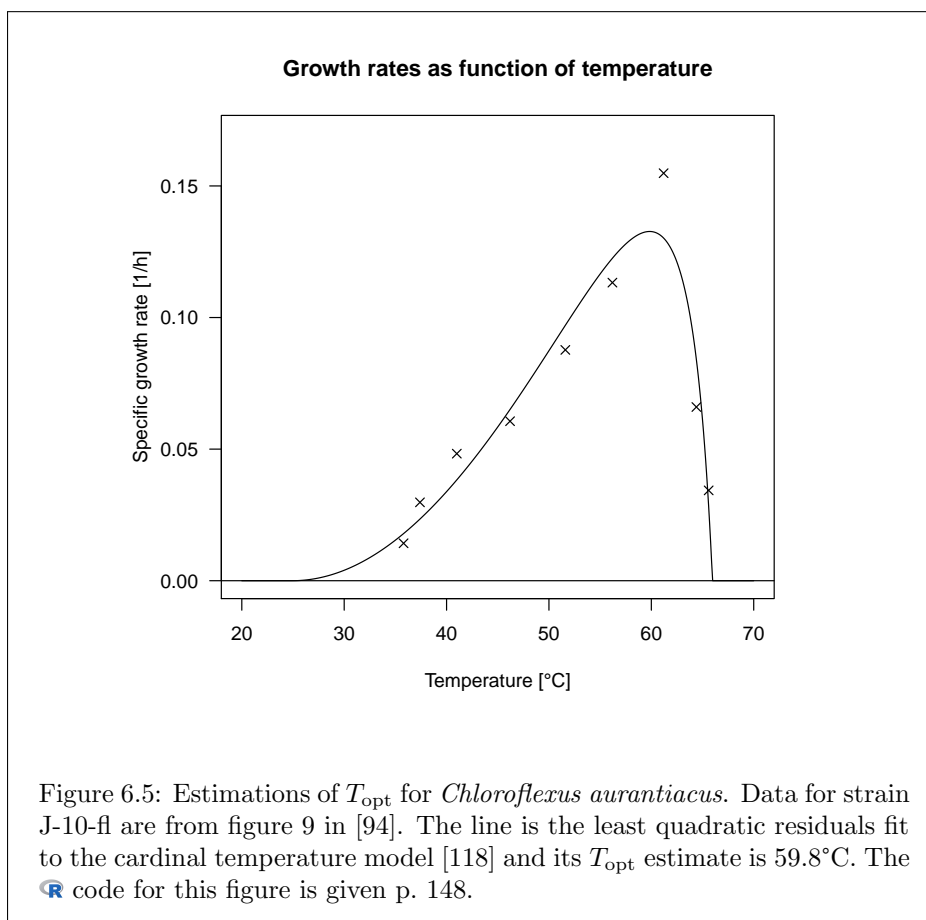
page 109. In [27], T_{opt} for strains MT4 and MT3 are reported as 87°C and 75°C, respectively. According to [161] “the isolates DSM 1616 and DSM 1617 of *Sulfolobus solfataricus* are probably identical with or similar to the “Caldariella” strains MT3 and MT4, isolated by de Rosa *et al.* (1975) [27].” Figure 6.4 page 110 gives 88.8°C and 77.4°C for strains MT4 and MT3, respectively. I have selected the value 79.9°C from the mean T_{opt} for the four strains in [45].

```
tocuT2[which(tocuT2$organism == "sulfolobus_solfataricus"), "topt"] <- 79.9
```



Species	Source 1	Source 2	Δ	T_{opt}
<i>Thermococcus kodakarensis</i>	95.0	60.0	35.0	85.0
<i>Synechococcus elongatus</i>	56.0	24.0	32.0	
<i>Shewanella violacea</i>	30.0	4.0	26.0	8.0
<i>Streptomyces glaucescens</i>	45.0	28.0	17.0	28.0
<i>Bacillus coagulans</i>	53.0	37.0	16.0	
<i>Sulfolobus solfataricus</i>	87.0	74.0	13.0	79.9
<i>Chloroflexus aurantiacus</i>	60.0	49.0	11.0	59.8
<i>Psychrobacter arcticus</i>	10.0	20.0	10.0	22.0
<i>Prochlorococcus marinus</i>	30.0	20.0	10.0	
<i>Nostoc punctiforme</i>	30.0	20.0	10.0	
<i>Streptomyces coelicolor</i>	37.0	28.0	9.0	28.0
<i>Helicobacter hepaticus</i>	37.0	46.0	9.0	37.0
<i>Streptococcus salivarius</i>	45.0	37.0	8.0	37.0
<i>Pyrococcus abyssi</i>	98.0	90.0	8.0	101.0
<i>Clostridium novyi</i>	45.0	37.0	8.0	45.0
<i>Bartonella bacilliformis</i>	37.0	29.0	8.0	28.0
<i>Hyphomonas neptunium</i>	33.5	26.0	7.5	37.0
<i>Asticcacaulis excentricus</i>	22.5	30.0	7.5	25.0
<i>Pseudomonas stutzeri</i>	37.0	30.0	7.0	35.0
<i>Nostoc</i> sp.	30.0	23.0	7.0	
<i>Burkholderia mallei</i>	30.0	37.0	7.0	27.2
<i>Bacillus anthracis</i>	30.0	37.0	7.0	39.5
<i>Synechococcus</i> sp.	27.5	21.0	6.5	
<i>Bacillus licheniformis</i>	38.5	32.0	6.5	50.0
<i>Haloferax volcanii</i>	42.3	36.0	6.3	45.3
<i>Vibrio parahaemolyticus</i>	37.0	31.0	6.0	36.0
<i>Thermotoga neapolitana</i>	85.0	79.0	6.0	77.4
<i>Pseudomonas pseudoalcaligenes</i>	35.0	29.0	6.0	
<i>Microcystis aeruginosa</i>	25.0	19.0	6.0	31.6
<i>Methanoplanus limicola</i>	40.0	34.0	6.0	40.0
<i>Synechocystis</i> sp.	28.5	23.0	5.5	32.0
<i>Desulfovibrio desulfuricans</i>	36.2	31.0	5.2	36.2
<i>Bacillus subtilis</i>	35.2	30.0	5.2	38.7

Table 6.1: Important T_{opt} discrepancies ($> 5^\circ\text{C}$) in decreasing order of magnitude (Δ). Source 1 is from [38, 73], Source 2 is from [31]. T_{opt} is the value used here as detailed in section 6.3.5 page 108. An empty value means that the corresponding entry was deleted in the final dataset.



FOR *Chloroflexus aurantiacus* according to the abstract from [103] T_{opt} is in the range from 52 to 60°C. According to table 1 from [36] T_{opt} is 55°C. According to [94] T_{opt} is 60°C and figure 6.5 page 112 is a re-creation of figure 9 from this paper. I have selected 59.8°C from figure 6.5 page 112.

```
tocuT2[which(tocuT2$organism == "chloroflexus_aurantiacus"), "topt"] <- 59.8
```

FOR *Psychrobacter arcticus* (273-4 = DSM 17307 = VKM B-2377) according to [6] “growth occurs at -10 to 28°C. Optimal growth temperature is 22°C.” I have selected this last value. *Planococcus halocryophilus* is able to grow at -15°C [88], liquid water is obtained with high solute concentrations so that cryophilic bacteria are also halophilic (tolerant to 19% NaCl in this case). It will be interesting to check if there is any convergent evolution with non-cryophilic halophilic bacteria. In [88] they used 5 statistics based on aminoacid frequencies defined in [4].

```
tocuT2[which(tocuT2$organism == "psychrobacter_arcticus"), "topt"] <- 22
```

FOR *Nostoc punctiforme* I found no data so I have deleted this entry for now.

Make forward biblio search for [4] and check for convergent evolution between non-cryophilic halophilic bacteria and cryophilic bacteria

Complete *Prochlorococcus marinus*

Check again for data for *Nostoc punctiforme*


```
todelete <- c(todelete, which(tocuT2$organism == "nostoc_punctiforme"))
```

FOR *Streptomyces coelicolor* according to chart 4 in [101] maximum yield is obtained at 30°C and the text state that “best growth and pigment intensity were observed at 28°C and 30°C”. I have selected 28°C for this species.

```
tocuT2[which(tocuT2$organism == "streptomyces_coelicolor"), "topt"] <- 28
```

FOR *Helicobacter hepaticus* according to [34] growth is observed at 37°C but not at 25°C and 42°C. I have selected 37°C for this species.

```
tocuT2[which(tocuT2$organism == "helicobacter_hepaticus"), "topt"] <- 37
```

FOR *Streptococcus salivarius* according to [120] for 290 strains “No growth takes place at 10°C nor at 47°C. The maximum temperature for growth is about 45°C, a minority of the cultures being able to grow at this temperature”. I have selected 37°C for this species.

```
tocuT2[which(tocuT2$organism == "streptococcus_salivarius"), "topt"] <- 37
```

FOR *Pyrococcus abyssi* (strain GE5 = CNCM I-1302) according to figure 3c from [32] T_{opt} is 96°C at atmospheric pressure. Figure 6.6 page 114 shows that T_{opt} is 101°C at *in situ* hydrostatic pressure. I have selected this last value.

```
tocuT2[which(tocuT2$organism == "pyrococcus_abyssi"), "topt"] <- 101
```

FOR *Clostridium novyi* I found no data, I have kept 45°C.

```
tocuT2[which(tocuT2$organism == "clostridium_novyi"), "topt"] <- 45
```

Check again *Clostridium novyi* for T_{opt} data

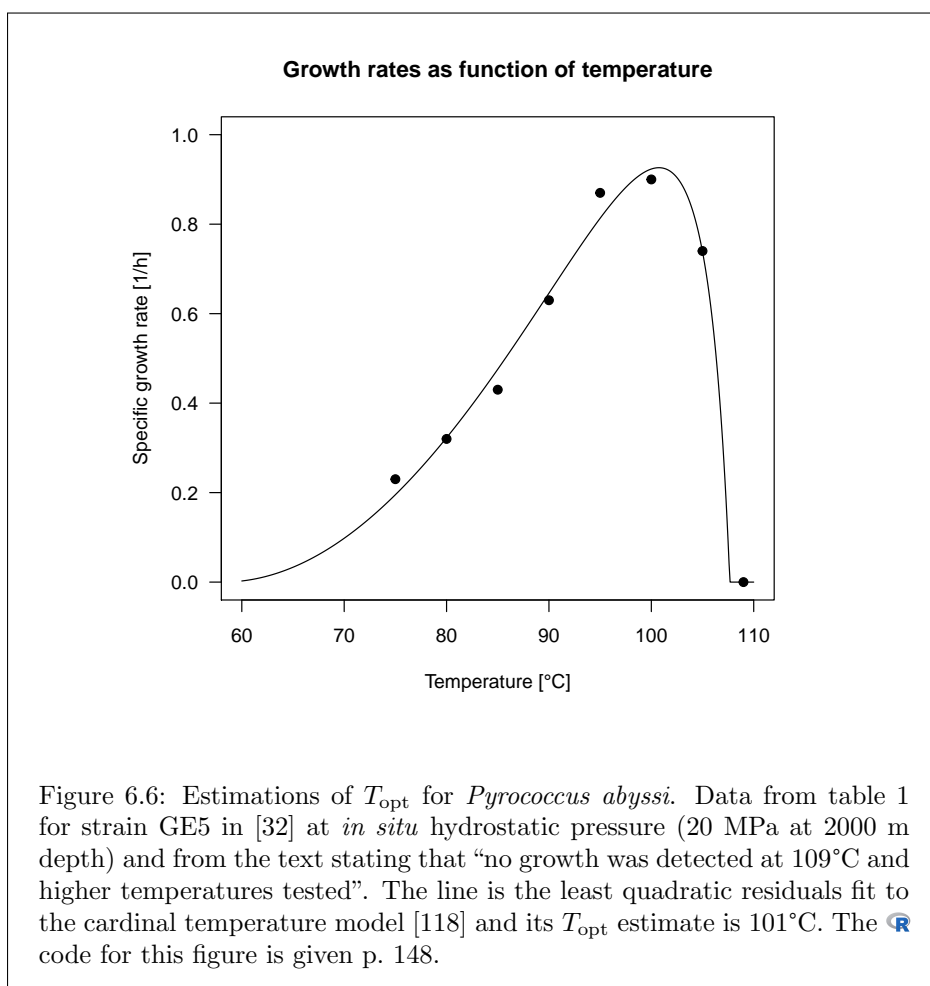
FOR *Bartonella bacilliformis* according to [81] “*Bartonella* spp. grow best in vitro at 37°C except for *Bartonella bacilliformis*, which grows best at 28°C.” I have kept 28°C.

```
tocuT2[which(tocuT2$organism == "bartonella_bacilliformis"), "topt"] <- 28
```

FOR *Hyphomonas neptunium* (LEIFSON 1964 LE670 = ATCC 15444 = IFAM LE6701) according to [85] the temperature range is from 4°C to 40°C with an optimum between 30°C and 37°C. Since the curve of the specific growth rate with respect to temperature is assymetric [118] I have kept 37°C.

```
tocuT2[which(tocuT2$organism == "hyphomonas_neptunium"), "topt"] <- 37
```

FOR *Asticcacaulis excentricus* according to table 2 from [149] T_{opt} is 30°C. These data are said to come from [107] in which there is a paragraph *Effect of Temperature on Growth* page 270 stating that: “Cultures of fresh-water, soil, and millipede isolates were routinely incubated at 30°C. Growth appears normal, but slower, at 25°C. The growth of two strains was tested at 37°C in agitated liquid cultures. The vibrioid strain, CB2, grew at the rate usually observed at 30°C, and the cells appeared normal. The growth of the second strain (bacteroid, CB11) was somewhat slower than at 30°C, as determined by turbidity measurements; most of the cells were elongated, and motile cells were absent. The marine isolates, which grew in enrichment cultures at 13°C and 19°C, grew more rapidly at 25°C. Growth was somewhat slower at 28°C than at 25°C”. The four *A. excentricus* strains are AC12, AC47, AC48, and KA4 according to page 292 in [107] and were all isolated from pond water according to table 1 page 236 and according to scheme 1 page 283 are all marine isolates. I have then used 25°C.



```
tocuT2[which(tocuT2$organism == "asticcacaulis_excentricus"), "topt"] <- 25
```

FOR *Pseudomonas stutzeri* according to [65] the temperature range is highly variable between strains, they wrote that: “The optimum temperature for growth is approximately 35°C” so I have have used this value.

```
tocuT2[which(tocuT2$organism == "pseudomonas_stutzeri"), "topt"] <- 35
```

FOR *Nostoc* sp. I have look for data when the species is documented and got nothing because I have previously deleted *N. punctiforme* and have no data for *N. azollae*. I have deleted *Nostoc* sp.

```
qui <- substr(tocuT2$organism, 1, 7) == "nostoc_"
tocuT2[qui, c("organism", "toptsp", "temperature")]
      organism toptsp temperature
1466  nostoc_azollae    NA         NA
1467 nostoc_punctiforme  30         20
1468  nostoc_sp        30         23
todelete <- c(todelete, which(tocuT2$organism == "nostoc_sp"))
```

FOR *Burkholderia mallei*, according to table 2 from [157] there is no growth at 41°C. *Burkholderia mallei* was peviously in the genus *Pseudomonas* homology group II, and from table 3 from [118] we have $T_{opt} = 27.2^\circ\text{C}$.

```
tocuT2[which(tocuT2$organism == "burkholderia_mallei"), "topt"] <- 27.2
```

FOR *Bacillus anthracis* according to figure 1 in [46] the temperature range is from 15°C to 45°C. According to table 2 in [51] the temperature range is from 10°C to 50°C and T_{opt} not reported in the literature. I found data in [143] and figure 6.7 page 116 shows why $T_{opt} = 39.5^\circ\text{C}$ was selected here.

```
tocuT2[which(tocuT2$organism == "bacillus_anthraxis"), "topt"] <- 39.5
```

FOR *Synechococcus* sp., this is a form-genus, that is an artificial group rather than a natural one. Figure 6.8 page 117 shows that T_{opt} is highly variable from strain to strain: the observed range from 46.8°C to 67°C is most likely underestimated here since there are no marine isolates in this sample. Figure 5.8 page 74 in [43] based on data from [104, 82] shows that T_{opt} ranges from 22°C to 67°C. I have therefore deleted this entry.

```
todelete <- c(todelete, which(tocuT2$organism == "synechococcus_sp"))
```

FOR *Bacillus licheniformis* according to table 2 from [150] T_{opt} is 49°C for one strain 51°C for the other one. I have then used 50°C.

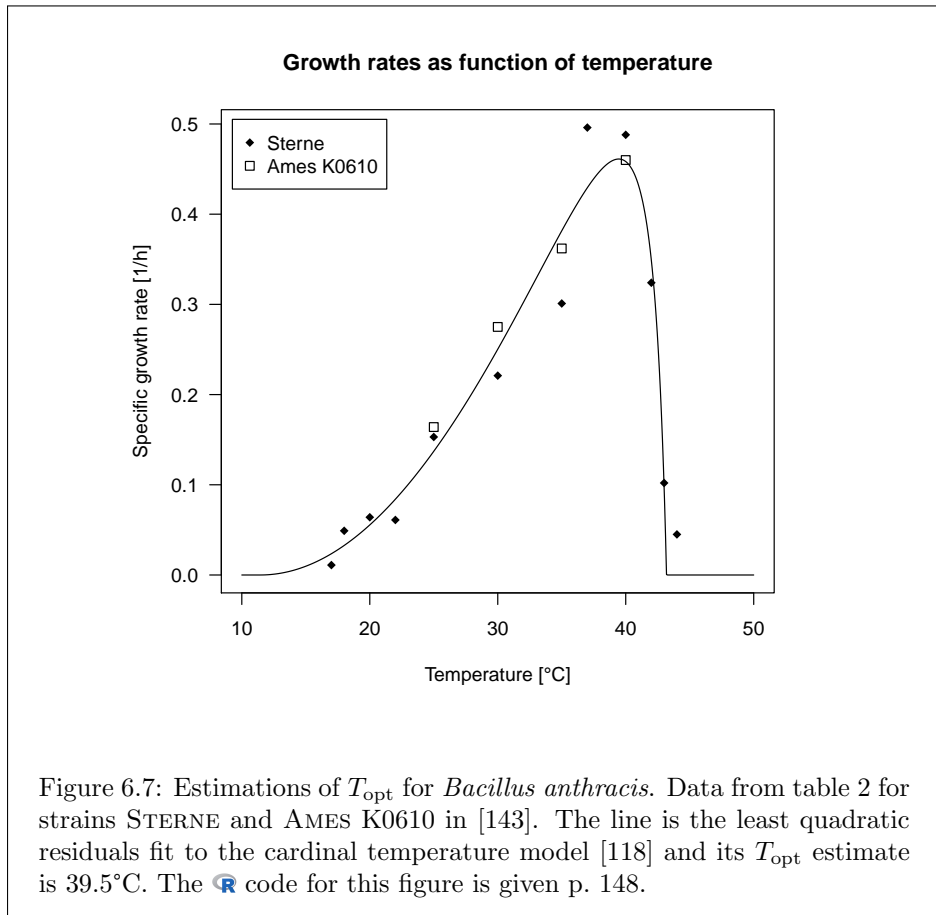
```
tocuT2[which(tocuT2$organism == "bacillus_licheniformis"), "topt"] <- 50
```

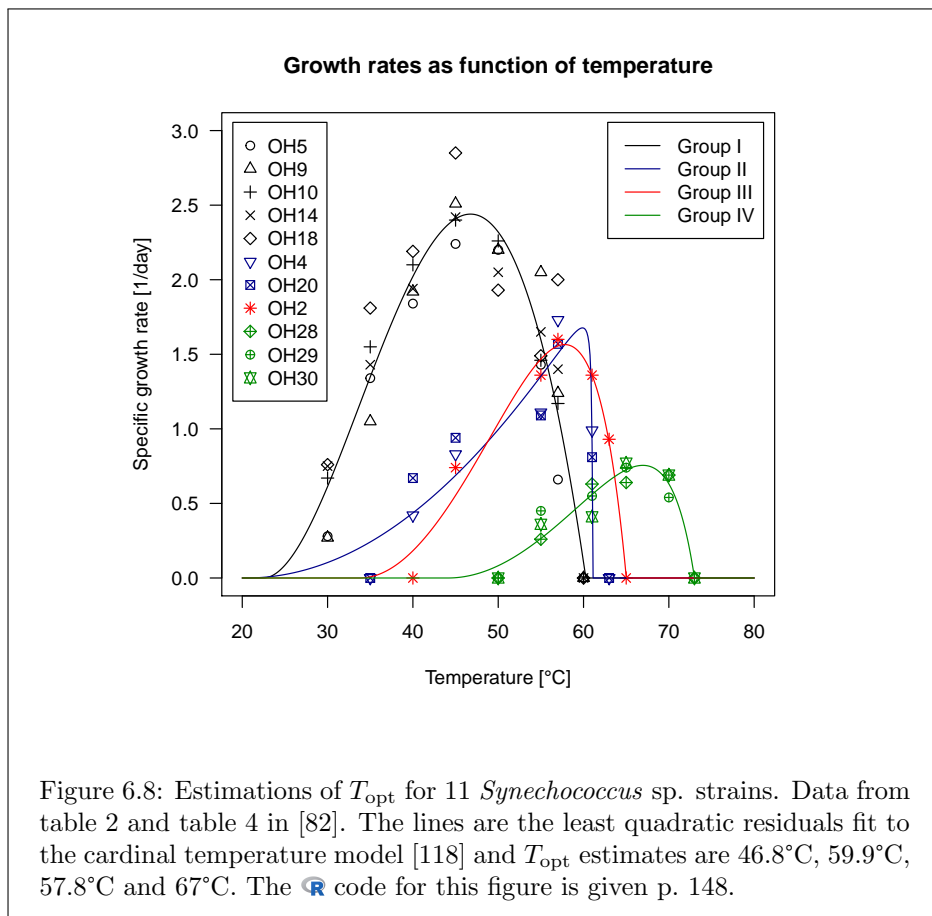
FOR *Haloferax volcanii* according to table 3 from [115] T_{opt} is 45°C and from reference 20 [95] in this table T_{opt} is 40°C. From figure 6.9 page 118 we have $T_{opt} = 45.3^\circ\text{C}$, I have then used this value.

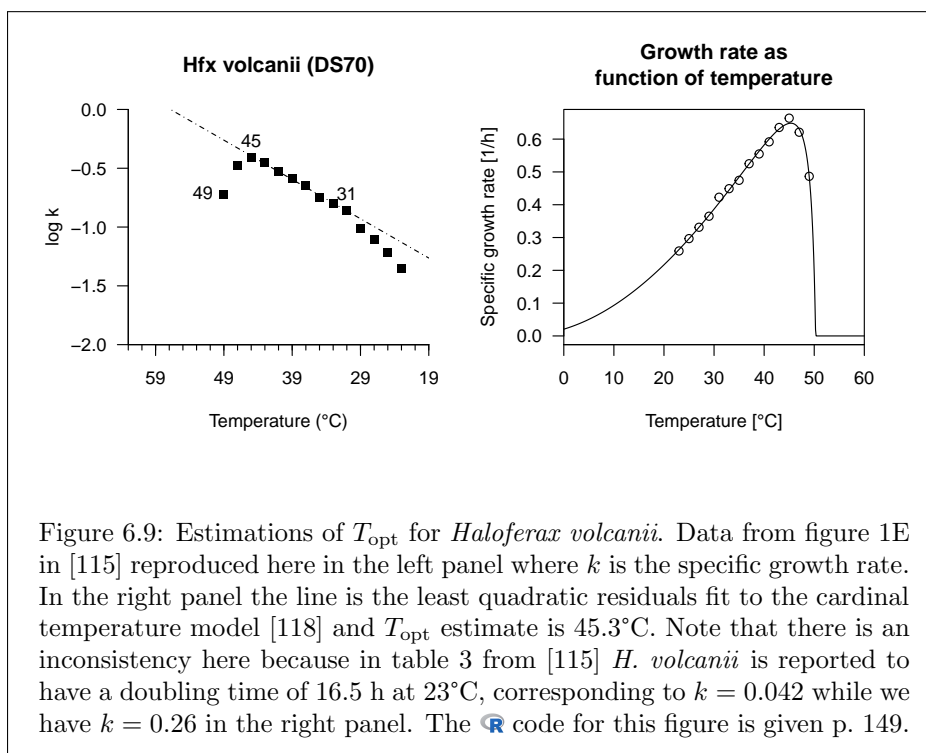
```
tocuT2[which(tocuT2$organism == "haloferax_volcanii"), "topt"] <- 45.3
```

FOR *Vibrio parahaemolyticus* according to [13] T_{opt} range is from 35°C to 37°C, I have then used 36°C.

```
tocuT2[which(tocuT2$organism == "vibrio_parahaemolyticus"), "topt"] <- 36
```







FOR *Thermotoga neapolitana* according to the introduction in [142]: “*T. maritima* and *T. neapolitana* grow optimally at 80°C and are hyperthermophilic species.” In [10] $T_{\text{opt}} = 77^\circ\text{C}$ is reported. From figure 6.10 page 119 we have $T_{\text{opt}} = 77.4^\circ\text{C}$, this last value was used.

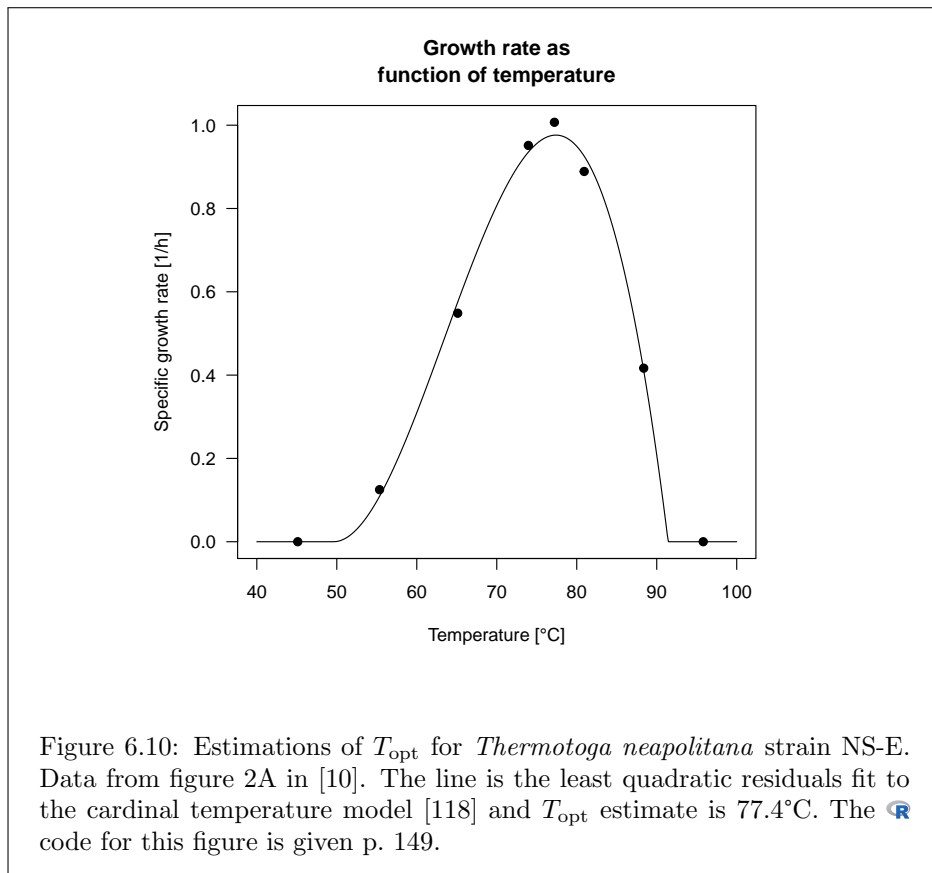
```
tocuT2[which(tocuT2$organism == "thermotoga_neapolitana"), "topt"] <- 77.4
```

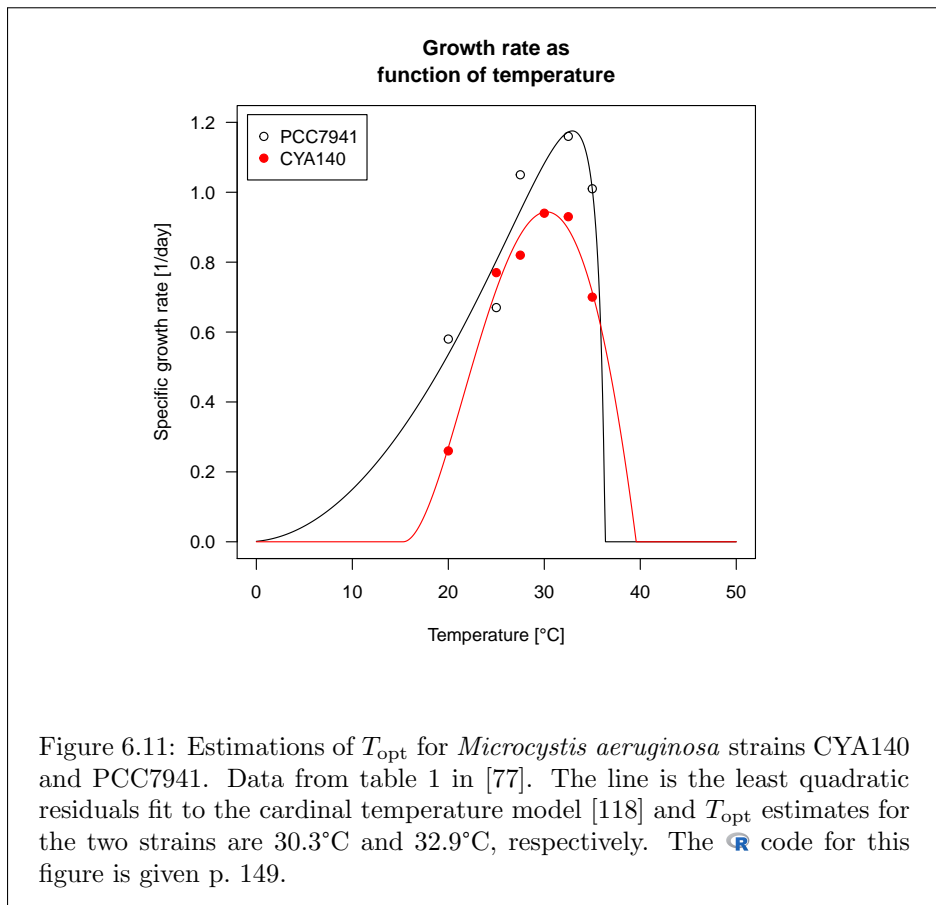
FOR *Pseudomonas pseudoalcaligenes* according to [126]: “Fig. 1 shows that the ancestral clone grew best at 35°C, but poorly at 45°C or higher temperature.” However, Fig. 1 in [126] represents the growth yield versus temperature, not the specific growth rate. There are only two points for the specific growth rate in table 3 from [126] which is not enough to fit the CTMI model. According to [121] *P. pseudoalcaligenes* should be reclassified as a later synonym of *Pseudomonas oleovorans*. *P. oleovorans* was first described in [68] but they just write that “It grows well either at room temperature (usually between 25 and 30°C) or at 37.5°C”. According to the abstract in [156]: “Optimum growth temperature and pH were 35°C and 8.0, respectively” but in the results section: “Growth was observed within the temperature range 20-45°C, while 30°C was the optimum temperature (data not shown)”. I have deleted this entry.

Check again for *Pseudomonas oleovorans*

```
todelete <- c(todelete, which(tocuT2$organism == "pseudomonas_pseudoalcaligenes"))
```

FOR *Microcystis aeruginosa* according to table 1 in [77] T_{opt} is 30°C for strain CYA140 and 32.5°C for strain PCC7941. From figure 6.11 page 120 we have 30.3 and 32.9°C, so I have used the mean 31.6°C here.






```
tocuT2[which(tocuT2$organism == "microcystis_aeruginosa"), "topt"] <- 31.6
```

FOR *Methanoplanus limicola* according to figure 1 in [153] T_{opt} is 40°C (there is no growth at 43°C). I have then used 40°C.

```
tocuT2[which(tocuT2$organism == "methanoplanus_limicola"), "topt"] <- 40
```

FOR *Synechocystis sp.* according to [158] T_{opt} is 32°C, I have then used this value.

```
tocuT2[which(tocuT2$organism == "synechocystis_sp"), "topt"] <- 32
```

FOR *Desulfovibrio desulfuricans* according to table 1 in [100] the maximal growth temperature is 40°C. I have then used the higher 36.25°C for T_{opt} .

```
tocuT2[which(tocuT2$organism == "desulfovibrio_desulfuricans"), "topt"] <- 36.25
```

FOR *Bacillus subtilis* in the paper for the CTMI model [118] we have an estimate of 38.7°C for T_{opt} from a dataset with 15 points from [110]. In [150] T_{opt} is 46°C for three strains (P, 168, B692) from *B. subtilis* but without detailed results. I have kept 38.7°C.

```
tocuT2[which(tocuT2$organism == "bacillus_subtilis"), "topt"] <- 38.7
```

THE following R code is used to generate the table summarizing the T_{opt} values used here. Column `toptsp` are data from [38, 73], column `temperature` data from [31] and column `topt` the value selected here.

```
both <- !(is.na(tocuT2$toptsp)) & !is.na(tocuT2$temperature)
qui <- tocuT2[both & tocuT2$delta > 5, c("organism", "toptsp", "temperature", "delta", "topt")]
head(qui[rev(order(qui$delta)), ])
      organism toptsp temperature delta topt
2137 thermococcus_kodakarensis    95         60    35 85.0
2085 synechococcus_elongatus     56         24    32  NA
1904 shewanella_violacea        30          4    26  8.0
2034 streptomyces_glaucescens    45         28    17 28.0
211  bacillus_coagulans         53         37    16  NA
2067 sulfolobus_solfataricus     87         74    13 79.9
```

NOW using the package `xtable`, it's easy to generate with the following code the L^AT_EX table 6.1 page 111 corresponding the the previous raw code output.

```
library(xtable)
mytab <- qui[rev(order(qui$delta)), ]
colnames(mytab) <- c("Species", "Source 1", "Source 2", "$\Delta", "\Topt{")
mytab[, 3] <- as.numeric(mytab[, 3])
mkspname <- function(x){
  tmp <- unlist(strsplit(x, split = "_"))
  substr(tmp[1], 1, 1) <- toupper(substr(tmp[1], 1, 1))
  if(tmp[2] != "sp")
    return(paste("\textit{" , tmp[1], tmp[2], "}")
  else
    return(paste("\textit{" , tmp[1], " } sp.")
}
mytab[, 1] <- sapply(mytab[, 1], mkspname)
caption <- "Important \Topt{ discrepancies (> 5°C) in decreasing order of
magnitude ($\Delta). Source 1 is from \cite{galtierlobry,LobryJR2006},
Source 2 is from \cite{EngqvistMKM2018}. \Topt{ is the value used here as
detailed in section \ref{solving} page \pageref{solving}. An
empty value means that the corresponding entry was deleted in the final
dataset."
print(xtable(mytab, caption = caption, digits = 1,
  label = "diffs"), file = "tables/diffs.tex", size = "normalsize",
  include.rownames = FALSE, sanitize.colnames.function = function(x){x},
  sanitize.text.function = function(x){x})
```

Tiens, il faudra que je vérifie qu'ils n'ont pas fait un plagiat par anticipation de [38]

THE following `R` code is to delete all the entries for which I was unable to find a reliable T_{opt} estimate in this section.

```
tocuT2 <- tocuT2[-todelete, ]
```

6.3.6 Collation finale des températures optimales de croissance

C'EST ici qu'il faut être particulièrement vigilant pour ne pas écraser les données villainement. Je veux un exemple de chaque cas de figure pour vérifier au fur et à mesure qu'il n'y a pas de problème. Pour résumer, j'ai la source de données [38, 73] et la source de données [31], donc 4 cas de figure possibles selon que la température optimale de croissance est documentée ou non dans chaque source. Mais dans le cas où la température est documentée dans les deux sources j'ai deux sous-cas : si l'écart est de plus de 5°C j'ai fait une recherche bibliographique pour résoudre le cas et sinon rien. Donc si je compte bien j'ai 5 cas de figure :

1. documenté ni dans [38, 73] ni dans [31];
2. documenté dans [38, 73] mais pas dans [31];
3. pas documenté dans [38, 73] mais dans [31];
4. documenté dans [38, 73] et dans [31] :
 - (a) écart important avec résolution manuelle ;
 - (b) écart peu important non examiné.

```
test <- with(tocuT2, which(is.na(toptsp) & is.na(temperature))[1])
test <- c(test, with(tocuT2, which(!is.na(toptsp) & is.na(temperature))[1]))
test <- c(test, with(tocuT2, which(is.na(toptsp) & !is.na(temperature))[1]))
test <- c(test, with(tocuT2, which(!is.na(toptsp) & !is.na(temperature) & !is.na(topt))[1]))
test <- c(test, with(tocuT2, which(!is.na(toptsp) & !is.na(temperature) & is.na(topt))[1]))
what <- c("organism", "toptsp", "temperature", "topt")
tocuT2[test, what]
      organism toptsp temperature topt
3  acetoanaerobium_sticklandii      NA          NA      NA
170 anaplasma_marginale      37.0          NA      NA
1  abiotrophia_defectiva      NA          37      NA
193 asticcacaulis_excentricus    22.5          30     25
4   acetobacter_pasteurianus    27.5          27      NA
```

JE compile maintenant les données étape par étape en vérifiant avec le jeu test que tout va bien.

```
tocuT2[is.na(tocuT2$topt), "topt"] <- tocuT2[is.na(tocuT2$topt), "toptsp"]
tocuT2[test, what]
      organism toptsp temperature topt
3  acetoanaerobium_sticklandii      NA          NA      NA
170 anaplasma_marginale      37.0          NA     37.0
1  abiotrophia_defectiva      NA          37      NA
193 asticcacaulis_excentricus    22.5          30     25.0
4   acetobacter_pasteurianus    27.5          27     27.5

tocuT2[is.na(tocuT2$topt), "topt"] <- tocuT2[is.na(tocuT2$topt), "temperature"]
tocuT2[test, what]
```

```

      organism toptsp temperature topt
3  acetoanaerobium_sticklandii      NA      NA      NA
170 anaplasma_marginale      37.0      NA      37.0
1    abiotrophia_defectiva      NA      37      37.0
193 asticcacaulis_excentricus     22.5     30      25.0
4    acetobacter_pasteurianus     27.5     27      27.5

sum(!is.na(tocuT2$topt)) # 1868
[1] 1869

```

J'ai donc l'usage du code et des données de température pour 1868 espèces bactériennes.

Intermediate backup, not usefull for final document

```
save(tocuT2, file = "local/tocuT2.Rda")
```

6.3.7 Manual bibliographical search for T_{opt}

Here is the list of species to be searched:

```

todosearch <- with(tocuT2, organism[!mined & !genuine])
# Todo: Remove obvious mesophilic species
boring <- c("borrelia", "chlamydia", "borreliella", "rickettsia", "wolbachia",
"mycobacterium", "pseudomonas", "agrobacterium", "helicobacter", "ehrlichia")
remove <- unlist(lapply(boring, grep, todosearch))
(todosearch[-remove])

[1] "acetoanaerobium_sticklandii"      "acetobacteraceae_bacterium"
[3] "acetomicrobium_hydrogeniformans"  "acholeplasma_NA"
[5] "acidaminococcus_sp"              "acidianus_hospitalis"
[7] "acidiphilium_sp"                 "acidocella_sp"
[9] "acidovorax_ebreus"               "aciduliprofundum_sp"
[11] "acinetobacter_oleivorans"         "actinobaculum_sp"
[13] "aggregatibacter_sp"              "alcanivorax_pacificus"
[15] "alkaliphilus_metalliredigens"     "alloprevotella_tannerae"
[17] "alpha_proteobacterium"           "alteromonas_naphthalenivorans"
[19] "amycolatopsis_methanolica"       "anaeromyxobacter_sp"
[21] "anaerostipes_sp"                 "anaerotruncus_sp"
[23] "anaplasma_centrale"              "archaeon_gw2011_ar10"
[25] "archaeon_gw2011_ar15"            "archaeon_gw2011_ar20"
[27] "arcobacter_sp"                   "aromatoleum_aromaticum"
[29] "atopobium_sp"                    "bacillus_paralicheniformis"
[31] "bacillus_toyonensis"             "bacterium_endosymbiont"
[33] "bacteroidales_bacterium"         "bacteroides_nordii"
[35] "bacteroidetes_oral"              "bartonella_sp"
[37] "baumannia_cicadellinicola"       "bernardetia_litoralis"
[39] "beta_proteobacterium"            "bilophila_sp"
[41] "blattabacterium_cuenoti"          "blattabacterium_sp"
[43] "blochmannia_endosymbiont"        "bradyrhizobiaceae_bacterium"
[45] "bradyrhizobium_diazoeficiens"    "brenneria_sp"
[47] "brucella_ceti"                   "brucella_microti"
[49] "brucella_pinnipedialis"          "burkholderiales_bacterium"
[51] "butyrate-producing_bacterium"    "campylobacter_peloridis"
[53] "campylobacter_subantarcticus"    "candidate_division"
[55] "cardinium_endosymbiont"          "cellulomonas_gilvus"
[57] "cellvibrio_sp"                   "chamaesiphon_minutus"
[59] "chelativorans_sp"                "chlamydophila_pecorum"
[61] "chlorobium_luteolum"             "chlorobium_phaeobacteroides"
[63] "chloroflexus_sp"                 "clostridiales_bacterium"
[65] "clostridioides_difficile"        "clostridium_cf"
[67] "coleofasciculus_chthonoplastes"  "collinsella_sp"
[69] "comamonadaceae_bacterium"        "coprobacillus_sp"
[71] "coprococcus_sp"                  "coriobacteriaceae_bacterium"
[73] "coxiella_endosymbiont"           "cyanobacterium_aponinum"
[75] "cyanobacterium_endosymbiont"     "cyanobacterium_stanieri"
[77] "cyanobium_gracile"               "cyanobium_sp"
[79] "cyanothecae_sp"                  "cycloclasticus_sp"
[81] "cycloclasticus_zanclis"          "cyndrospermum_stagnale"
[83] "dactylococcopsis_salina"         "dehalobacter_sp"

```

[85]	"dehalococcoides_mccartyi"	"dehalogenimonas_lykanthroporepellens"
[87]	"deinococcus_swuensis"	"dermabacter_sp"
[89]	"desmospora_sp"	"desulfitobacterium_dichloroeliminans"
[91]	"desulfococcus_oleovorans"	"dialister_succinatiphilus"
[93]	"dokdonia_sp"	"dorea_sp"
[95]	"edwardsiella_anguillarum"	"eggerthella_sp"
[97]	"endosymbiont_of"	"enterobacter_lignolyticus"
[99]	"erysipelatoclostridium_amosum"	"erysipelotrichaceae_bacterium"
[101]	"erythrobacter_sp"	"eubacterium_plexicaudatum"
[103]	"faecalibacterium_cf"	"fermentimonas_caenicola"
[105]	"ferroplasma_acidarmanus"	"fimbriimonas_ginsengisoli"
[107]	"firmicutes_bacterium"	"flavobacteria_bacterium"
[109]	"flavobacteriaceae_bacterium"	"francisella_cf"
[111]	"francisella_sp"	"frankia_casuarinae"
[113]	"frankia_inefficax"	"frankia_symbiont"
[115]	"gallionella_capsiferriiformans"	"gamma_proteobacterium"
[117]	"geitlerinema_sp"	"gemmatirosa_kalamazooensis"
[119]	"geobacillus_genomosp"	"geobacter_uraniireducens"
[121]	"glaciecola_nitratireducens"	"gloeobacter_kilauensis"
[123]	"gloeocapsa_sp"	"gynuella_sunshinyii"
[125]	"halanaeroarchaeum_sulfuriireducens"	"halanaerobium_hydrogeniformans"
[127]	"halonotius_sp"	"halophilic_archaeon"
[129]	"haloquadratum_sp"	"halothece_sp"
[131]	"halothermothrix_orenii"	"hydrogenobaculum_sp"
[133]	"hymenobacter_swuensis"	"ilumatobacter_coccineus"
[135]	"jannaschia_sp"	"janthinobacterium_sp"
[137]	"jeotgalibacillus_malaysiensis"	"jeotgalicoccus_saudimassiliensis"
[139]	"kangiella_geojedonensis"	"ketogulonicigenium_vulgare"
[141]	"kitasatospora_cheerisanensis"	"komagataeibacter_medellinensis"
[143]	"kordia_algicida"	"kutzneria_sp"
[145]	"lachnoanaerobaculum_sp"	"lachnoclostridium_phytofermentans"
[147]	"lachnospiraceae_oral"	"lawsonia_intracellularis"
[149]	"leeuwenhoekella_blandensis"	"legionella_drancourtii"
[151]	"leptothrix_ochracea"	"leptotrichia_sp"
[153]	"leuconostoc_kimchii"	"lysiniibacillus_saudimassiliensis"
[155]	"lysiniibacillus_varians"	"mageeibacillus_indolicus"
[157]	"magnetococcus_marinus"	"magnetospira_sp"
[159]	"mannheimia_sp"	"maribacter_sp"
[161]	"marine_actinobacterium"	"marine_gamma"
[163]	"marinobacter_nanhaiticus"	"marinomonas_posidonica"
[165]	"martellella_endophytica"	"mesorhizobium_australicum"
[167]	"mesorhizobium_opportunatum"	"metallophaera_cuprina"
[169]	"metallophaera_yellowstonensis"	"methanocaldococcus_sp"
[171]	"methanolobus_psychrophilus"	"methanoseta_concillii"
[173]	"methanoseta_thermophila"	"methylophilum_inferorum"
[175]	"methylobacterium_nodulans"	"methylobacterium_populi"
[177]	"methylophilaceae_bacterium"	"methylovorus_glucosetrophus"
[179]	"methylovorus_sp"	"micavibrio_aeruginosavorus"
[181]	"microvirga_lotononidis"	"mitsuokella_sp"
[183]	"moorea_producens"	"mucinivorans_hirudinis"
[185]	"mucispirillum_schaedleri"	"muricauda_lutaonensis"
[187]	"myroides_sp"	"nitratiruptor_sp"
[189]	"nitrosococcus_halophilus"	"nitrosococcus_watsonii"
[191]	"nitrosomonas_sp"	"nitrosopumilus_maritimus"
[193]	"nitrospira_defluvii"	"nocardiodaceae_bacterium"
[195]	"nonlabens_marinus"	"nostoc_azollae"
[197]	"oceanicaulis_sp"	"oceanimonas_sp"
[199]	"oleiagrionas_soli"	"olsenella_sp"
[201]	"opitutaceae_bacterium"	"oribacterium_sp"
[203]	"oscillatoria_nigro-viridis"	"oscillatoriales_cyanobacterium"
[205]	"oscillibacter_sp"	"oscillochloris_trichoides"
[207]	"paenisporosarcina_sp"	"pandoraea_sp"
[209]	"parabacteroides_sp"	"parageobacillus_genomosp"
[211]	"parageobacillus_thermoglucosidasius"	"pectobacterium_parmentieri"
[213]	"pectobacterium_sp"	"pelodictyon_phaeoclathratiforme"
[215]	"pelosinus_sp"	"peptoanaerobacter_stomatis"
[217]	"peptoclostridium_acidaminophilum"	"peptoniphilus_sp"
[219]	"peptostreptococcaceae_bacterium"	"phaeobacter_sp"
[221]	"phascolarctobacterium_sp"	"phascolarctobacterium_succinatutens"
[223]	"photobacterium_sp"	"phycisphaera_mikurensis"
[225]	"plautia_stali"	"polaribacter_sp"
[227]	"polymorphum_gilvum"	"polynucleobacter_asymbioticus"

[229]	"pontibacter_korlensis"	"porphyromonas_sp"
[231]	"prevotella_sp"	"prochlorococcus_sp"
[233]	"propionibacterium_humerusii"	"prosthecochloris_aestuarii"
[235]	"pseudarthrobacter_phenanthrenivorans"	"pseudothermotoga_thermarum"
[237]	"pseudovibrio_sp"	"pusillimonas_sp"
[239]	"pyrococcus_yayanosii"	"rhizobium_gallicum"
[241]	"rhodobacteraceae_bacterium"	"rhodobacterales_bacterium"
[243]	"rhodoluna_lacicola"	"rivularia_sp"
[245]	"roseibium_sp"	"roseiflexus_sp"
[247]	"rugosibacter_aromaticivorans"	"ruminiclostridium_thermocellum"
[249]	"ruminococcaceae_bacterium"	"ruminococcus_bicirculans"
[251]	"salinarchaeum_sp"	"sar86_cluster"
[253]	"secondary_endosymbiont"	"sediminispirochaeta_smaragdinae"
[255]	"selenomonas_sp"	"shewanella_piezotolerans"
[257]	"siansivirga_zeaxanthinifaciens"	"silicibacter_lacuscaerulensis"
[259]	"silicibacter_sp"	"sneathia_amnii"
[261]	"sodalis_praeaptivus"	"sphingomonas_hengshuiensis"
[263]	"sphingomonas_taxi"	"sphingopyxis_fribergensis"
[265]	"spiribacter_curvatus"	"spiribacter_salinus"
[267]	"spirosoma_radiotolerans"	"stanieria_cyanosphaera"
[269]	"strawberry_lethal"	"streptomyces_bingchenggensis"
[271]	"streptomyces_davawensis"	"streptomyces_pratensis"
[273]	"streptomyces_roseosporus"	"streptomycetaceae_bacterium"
[275]	"subdoligranulum_sp"	"sulfuricella_sp"
[277]	"sulfurihydrogenibium_sp"	"sulfurovum_sp"
[279]	"synergistes_sp"	"synthetic_escherichia"
[281]	"tannerella_sp"	"thalassobium_sp"
[283]	"thermincola_potens"	"thermococcus_eurythermalis"
[285]	"thermococcus_nautili"	"thermococcus_onnurineus"
[287]	"thermofilum_adornatus"	"thermofilum_carboxyditrophus"
[289]	"thermofilum_uzonense"	"thermogladus_cellulolyticus"
[291]	"thermoplasmatales_archaeon"	"thermosynechococcus_sp"
[293]	"thioalkalivibrio_sp"	"thioalkalivibrio_sulfidiphilus"
[295]	"thioflavococcus_mobilis"	"thiomonas_sp"
[297]	"thioploca_ingrica"	"thiorhodovibrio_sp"
[299]	"treponema_paraluiscuniculi"	"trichodesmium_erythraeum"
[301]	"tyzzerella_nexilis"	"uncultured_termite"
[303]	"veillonella_sp"	"verrucomicrobia_bacterium"
[305]	"verrucomicrobiae_bacterium"	"vulcanisaeta_moutnovskia"
[307]	"winogradskyella_sp"	"xanthomonas_cannabis"

Clostridium sticklandii was recently (2016) reclassified as *Acetoanaerobium sticklandii* [37]. According to [37]: “The description of *Acetoanaerobium sticklandii* is identical to that provided earlier for *Clostridium sticklandii* (STADTMAN & McCLUNG, 1957 [132]; RAINEY *et al.*, 2009 [113]). The type strain is ATCC 12662^T = DSM 519^T = JCM 1433^T, isolated from black mud from the east shore of San Francisco Bay (STADTMAN & BARKER, 1951 [131]).” From the description in [132] we have: “Grows well from 30 to 38 °C”. I don’t have access to [113]. There are no T_{opt} data in [131]. So the best guess I have is that it’s a mesophilic species but no more.

Halanaeroarchaeum sulfurireducens was recently (2016) described [129] as the “first obligately anaerobic sulfur-respiring haloarchaeon”. The description in [129] states that: “The optimum growth temperature is 37–40 °C (maximum 46°C)”. They also write in [129] page 2380 that “the details of anaerobic growth kinetics have been described previously (SOROKIN *et al.*, 2016 [128])”. In [128] is written page 245: “with the optimum growth temperature of 40°C”. I will stick with this last value.

Mesophilic species are boring, try to enrich first in extremophiles

```
i <- which(tocuT2$organism == "halanaeroarchaeum_sulfurireducens")
tocuT2[i, "topt"] <- 40
tocuT2[i, "family"] <- 2236 # Halobacteriaceae
tocuT2[i, "order"] <- 2235 # Halobacteriales
tocuT2[i, "class"] <- 183963 # Halobacteria
```

```

tocuT2[i, "phylum"] <- 28890 # Euryarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea

```

Halanaerobium hydrogeniformans synonyms are *Halanaerobium sapolanicus* and *Halanaerobium* sp. ‘sapolanicus’. According to [119] *H. hydrogeniformans* has not been validly published yet. According to [18] T_{opt} is 33°C.

```

i <- which(tocuT2$organism == "halanaerobium_hydrogeniformans")
tocuT2[i, "topt"] <- 33
tocuT2[i, "family"] <- 972 # Halanaerobiaceae
tocuT2[i, "order"] <- 53433 # Halanaerobiales
tocuT2[i, "class"] <- 186801 # Clostridia
tocuT2[i, "phylum"] <- 1239 # Firmicutes
tocuT2[i, "superkingdom"] <- 2 # Bacteria

```

Halonotius is a new genus described in 2010 [20] with a single species *H. pteroides*. According to [20]: “The optimum temperature for growth was 37-40°C, depending on the strain, and no growth was observed at 4 or 55°C”.

```

i <- which(tocuT2$organism == "halonotius_sp")
tocuT2[i, "topt"] <- 39
tocuT2[i, "family"] <- 1963271 # Halorubraceae
tocuT2[i, "order"] <- 1644055 # Haloferacales
tocuT2[i, "class"] <- 183963 # Halobacteria
tocuT2[i, "phylum"] <- 28890 # Euryarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea

```

FOR halophilic_archaeon I need to dig back since this is obviously not a regular taxonomic name.

```

load("local/bact2.Rda")
bact[grep("HALOPHILIC ARCHAEON", bact$species), c("species", "nCDS")]

```

	species	nCDS
50	HALOPHILIC ARCHAEON J07HX5	2128
51	HALOPHILIC ARCHAEON J07HX64	3026
52	HALOPHILIC ARCHAEON J07HB67	2833
241	HALOPHILIC ARCHAEON DL31	3483

We have then data merged for the complete genome of 4 strains of an “halophilic archaeon”. Typical metagenomic data, no way to get T_{opt} from this.

FOR haloquadratum_sp I need again to dig a little back: the entry is then for *Haloquadratum* sp. strain J07HQX50 from [106]. According to [147] this strain is “the first genome of a separate candidate species of the genus *Haloquadratum* (J07HQX50) [106]”. These are metagenomic data but I love *Haloquadratum* too much to delete this entry, I will use *Haloquadratum walsbyi* as a proxy for T_{opt} .

```

load("local/bact2.Rda")
bact[grep("HALOQUADRATUM SP", bact$species), c("species", "nCDS")]

```

	species	nCDS
58	HALOQUADRATUM SP. J07HQX50	2866

```

proxy <- which(tocuT2$organism == "haloquadratum_walsbyi")
i <- which(tocuT2$organism == "haloquadratum_sp")
tocuT2[i, "topt"] <- tocuT2[proxy, "topt"] # 37
tocuT2[i, "family"] <- tocuT2[proxy, "family"]
tocuT2[i, "order"] <- tocuT2[proxy, "order"]
tocuT2[i, "class"] <- tocuT2[proxy, "class"]
tocuT2[i, "phylum"] <- tocuT2[proxy, "phylum"]
tocuT2[i, "superkingdom"] <- tocuT2[proxy, "superkingdom"]

```

Halothece is a new genus described in 2008 [112] that should be corrected to *Halothece* to follow both the Botanical Code and the Bacteriological Code rules [96]. According to [8]: “Extremely halophilic *Aphanothece* spp.

(=*Halotheca*) from a solar pond near the Dead Sea grew at 48°C but not at 50°C (DOR and HORNOFF 1985) [7].”.

```
i <- which(tocu2$organism == "halotheca_sp")
tocu2[i, "topt"] <- 48
tocu2[i, "family"] <- 1890450 # Aphanothecaceae
tocu2[i, "order"] <- 1118 # Chroococcales
tocu2[i, "class"] <- NA
tocu2[i, "phylum"] <- 1117 # Cyanobacteria
tocu2[i, "superkingdom"] <- 2 # Bacteria
```

Halothermothrix orenii was described in 1994 [22]. They state in the description section that: “The optimum temperature for growth is 60°C.”

```
i <- which(tocu2$organism == "halothermothrix_oreonii")
tocu2[i, "topt"] <- 60
tocu2[i, "family"] <- 972 # Halanaerobiaceae
tocu2[i, "order"] <- 53433 # Halanaerobiales
tocu2[i, "class"] <- 186801 # Clostridia
tocu2[i, "phylum"] <- 1239 # Firmicutes
tocu2[i, "superkingdom"] <- 2 # Bacteria
```

Thermococcus eurythermalis was described in 2015 [159]. They state that: “Growth occurs over the temperature range 50-100°C (optimal growth at 85°C) at 0.1 MPa and extends to 102°C at 10 MPa.” The type strain A501^T was isolated from a hydrothermal vent site at a depth of 2 km, so 20 MPa for *in situ* pressure. From figure 2 in [159] we can see that at 20 MPa the specific growth rate is higher at 85°C than 95°C.

```
i <- which(tocu2$organism == "thermococcus_eurythermalis")
tocu2[i, "topt"] <- 85
tocu2[i, "family"] <- 2259 # Thermococcaceae
tocu2[i, "order"] <- 2258 # Thermococcales
tocu2[i, "class"] <- 183968 # Thermococci
tocu2[i, "phylum"] <- 28890 # Euryarchaeota
tocu2[i, "superkingdom"] <- 2157 # Archaea
```

Thermococcus nautili was described in 2014 [39] with strain 30-1^T as type strain. The description section states that: “Optimal growth occurs at 87.5°C (range 55– 90°C)”.

```
i <- which(tocu2$organism == "thermococcus_nautili")
tocu2[i, "topt"] <- 87.5
tocu2[i, "family"] <- 2259 # Thermococcaceae
tocu2[i, "order"] <- 2258 # Thermococcales
tocu2[i, "class"] <- 183968 # Thermococci
tocu2[i, "phylum"] <- 28890 # Euryarchaeota
tocu2[i, "superkingdom"] <- 2157 # Archaea
```

Thermococcus onnurineus was described in 2006 [5] with type strain NA1^T (=KCTC 10859^T, =JCM 13517^T). The description section states that: “Growth occurs at 63-90°C, with the optimum at 80°C”.

```
i <- which(tocu2$organism == "thermococcus_onnurineus")
tocu2[i, "topt"] <- 80
tocu2[i, "family"] <- 2259 # Thermococcaceae
tocu2[i, "order"] <- 2258 # Thermococcales
tocu2[i, "class"] <- 183968 # Thermococci
tocu2[i, "phylum"] <- 28890 # Euryarchaeota
tocu2[i, "superkingdom"] <- 2157 # Archaea
```

Thermofilum adornatus strain 1910b^T complete genome sequence was announced in 2013 [30]. The paper states: “The strain grows optimally at 92°C”.

```
i <- which(tocuT2$organism == "thermofilum_adornatus")
tocuT2[i, "topt"] <- 92
tocuT2[i, "family"] <- 114378 # Thermofilaceae
tocuT2[i, "order"] <- 2266 # Thermoproteales
tocuT2[i, "class"] <- 183924 # Thermoprotei
tocuT2[i, "phylum"] <- 28889 # Crenarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea
```

Thermofilum carboxyditrophus in table 3.2 page 13 in [144] T_{opt} is given as 90°C with a reference to [59].

```
i <- which(tocuT2$organism == "thermofilum_carboxyditrophus")
tocuT2[i, "topt"] <- 90
tocuT2[i, "family"] <- 114378 # Thermofilaceae
tocuT2[i, "order"] <- 2266 # Thermoproteales
tocuT2[i, "class"] <- 183924 # Thermoprotei
tocuT2[i, "phylum"] <- 28889 # Crenarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea
```

Check data in table 3.2
in [144]

Thermofilum uzonense type strain 1807-2^T (= DSM 28062^T = JCM 19810^T) was described in 2015 [146]. They describe the type strain as a: “Hyper-thermophile growing optimally at 85°C”.

```
i <- which(tocuT2$organism == "thermofilum_uzonense")
tocuT2[i, "topt"] <- 85
tocuT2[i, "family"] <- 114378 # Thermofilaceae
tocuT2[i, "order"] <- 2266 # Thermoproteales
tocuT2[i, "class"] <- 183924 # Thermoprotei
tocuT2[i, "phylum"] <- 28889 # Crenarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea
```

Thermogladius cellulolyticus strain 1633 complete genome sequence was announced in 2012 [80]. They wrote that: “Strain 1633 is an obligate anaerobe growing optimally at a temperature of 84°C”.

```
i <- which(tocuT2$organism == "thermogladius_cellulolyticus")
tocuT2[i, "topt"] <- 84
tocuT2[i, "family"] <- 2271 # Desulfurococcaceae
tocuT2[i, "order"] <- 114380 # Desulfurococcales
tocuT2[i, "class"] <- 183924 # Thermoprotei
tocuT2[i, "phylum"] <- 28889 # Crenarchaeota
tocuT2[i, "superkingdom"] <- 2157 # Archaea
```

```
load("local/bact2.Rda")
bact[grep("THERMOPLASMATALES", bact$species), c("species", "nCDS")]
      species nCDS
146 THERMOPLASMATALES ARCHAEON BRNA1 1528
```

Thermoplasmatales archaeon strain BRNA1 is an unclassified Thermoplasmatales. I’m giving up on this one because I’m aggregating data at the species level so that the information just on the family is not accurate enough.

```
load("local/bact2.Rda")
bact[grep("THERMOSYNECHOCOCCUS", bact$species), c("species", "nCDS")]
      species nCDS
12237 THERMOSYNECHOCOCCUS SP. NK55A 2233
```

Thermosynechococcus sp. strain NK55a complete genome sequence was announced in 2014 [134]. They wrote that: “The cyanobacterium *Thermosynechococcus* sp. strain NK55a (NBRC 108920) was isolated from a green microbial mat at the Nakabusa hot spring, Nagano Prefecture, Japan. NK55a and related strains are the major oxygenic photosynthetic organisms in mats growing at moderate temperatures of 52 to 60°C [33]”.


```
i <- which(tocuT2$organism == "thermosynechococcus_sp")
tocuT2[i, "topt"] <- 56
tocuT2[i, "family"] <- 1890426
tocuT2[i, "order"] <- 1890424
tocuT2[i, "class"] <- NA
tocuT2[i, "phylum"] <- 1117 # Cyanobacteria
tocuT2[i, "superkingdom"] <- 2 # Eubacteria
```

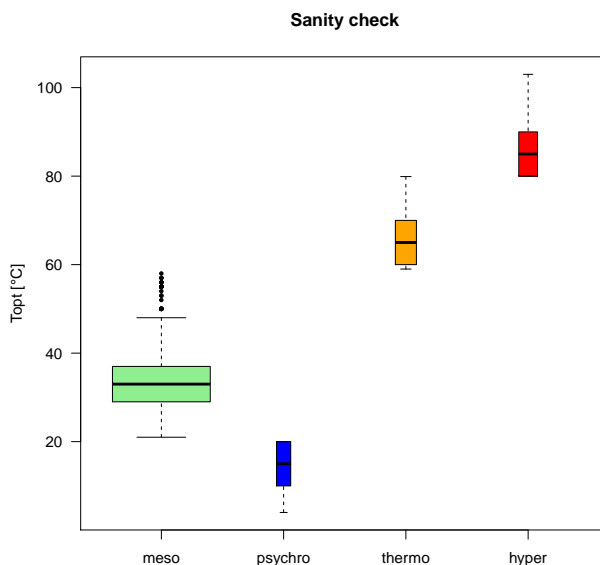
6.4 Polishing data

6.4.1 Sélection des lignes et colonnes, tri et sauvegarde

JE veux ne conserver que les espèces pour lesquelles la température est documentée, calculer les classes de thermophilie, trier pour que les mésophiles soient en dessous dans les graphiques et sauvegarder le tout.

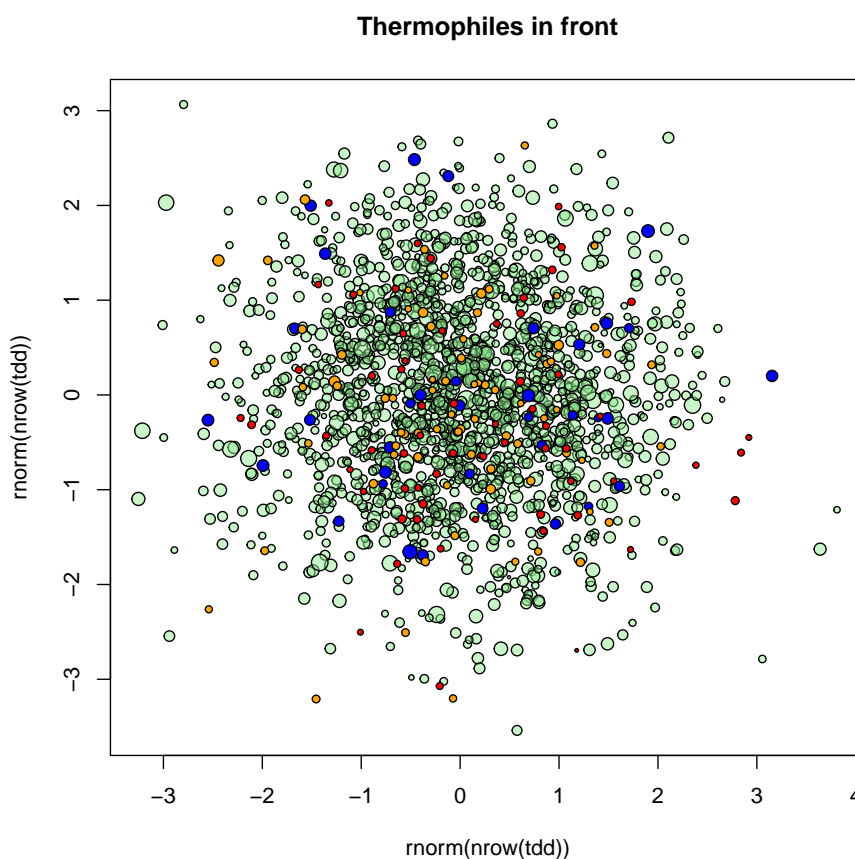
```
tdd <- tocuT2[!is.na(tocuT2$topt), ]
tdd <- tdd[ , -which(names(tdd) %in% c("toptsp", "temperature"))]
tdd$thermoclass <- with(tdd, ifelse(topt <= 20, "psychro", NA))
tdd$thermoclass <- with(tdd, ifelse(is.na(thermoclass) & topt < 59, "meso", thermoclass))
tdd$thermoclass <- with(tdd, ifelse(is.na(thermoclass) & topt < 80, "thermo", thermoclass))
tdd$thermoclass <- with(tdd, ifelse(is.na(thermoclass) & topt >= 80, "hyper", thermoclass))
table(tdd$thermoclass, useNA = "always")
  hyper    meso psychro  thermo    <NA>
    63    1705     36     79         0

tdd$thermoclass <- factor(tdd$thermoclass,
  levels = c("meso", "psychro", "thermo", "hyper"), ordered = TRUE)
mycols <- c("palegreen2", "blue", "orange", "red")
tdd$thermocols <- mycols[tdd$thermoclass]
boxplot(tdd$topt~tdd$thermoclass, col = mycols, pch = 19, cex = 0.5, varwidth = TRUE,
  main = "Sanity check", ylab = "Topt [°C]", las = 1)
getgenre <- function(x){
  res <- unlist(strsplit(x, split = "_"))[1]
  substr(res, 1, 1) <- toupper(substr(x, 1, 1))
  return(res)
}
tdd$genre <- sapply(tdd$organism, getgenre)
```



TRIONS maintenant et faisons un test de représentation graphique. Je ne garde une couleur transparente que pour les mésophiles.

```
tdd <- tdd[order(tdd$thermoclass), ]
library(sequinr)
tdd$athermocols <- ifelse(tdd$thermocols == "palegreen2", col2alpha("palegreen2", 0.5), tdd$thermocols)
tdd$cex <- sqrt(rowSums(tdd[, 2:62]/mean(rowSums(tdd[, 2:62])))
set.seed(1)
plot(rnorm(nrow(tdd)), rnorm(nrow(tdd)), pch = 21, bg = tdd$athermocols,
      cex = tdd$cex, main = "Thermophiles in front")
```



IL va me manquer des informations taxonomiques pour les espèces non documentées dans MKME. Je me base sur le genre qui est toujours documenté pour boucher les trous automatiquement.

```
tdd[is.na(tdd$domain), 1]
[1] "anaplasma_marginale"           "anaplasma_phagocytophilum"
[3] "asticcacaulis_biprosthecum"   "borrelia_afzelii"
[5] "borrelia_burgdorferi"         "borrelia_garinii"
[7] "buchnera_aphidicola"          "caulobacter_crescentus"
[9] "chlamydomphila_pneumoniae"    "chlorobium_chlorochromatii"
[11] "chloroherpeton_thalassium"     "coxiella_burnetii"
[13] "ehrlichia_canis"               "ehrlichia_chaffeensis"
[15] "ehrlichia_ruminantium"        "hahella_chejuensis"
[17] "halanaeroarchaeum_sulfurireducens" "halanaerobium_hydrogeniformans"
[19] "halonotius_sp"                 "haloquadratum_sp"
[21] "halothece_sp"                  "mannheimia_succiniciproducens"
[23] "mycobacterium_leprae"          "nitrospina_gracilis"
```

```

[25] "nodularia_spumigena"           "orientia_tsutsugamushi"
[27] "propionibacterium_acnes"       "ralstonia_eutropha"
[29] "rhodococcus_equi"             "rickettsia_conorii"
[31] "rickettsia_felis"             "rickettsia_montanensis"
[33] "rickettsia_prowazekii"        "rickettsia_rickettsii"
[35] "rickettsia_typhi"             "ruegeria_sp"
[37] "thermosynechococcus_sp"       "tolypothrix_sp"
[39] "tropheryma_whipplei"         "vibrio_fischeri"
[41] "wigglesworthia_glossinidia"   "wolbachia_endosymbiont"
[43] "cenarchaeum_symbiosum"        "halothermothrix_oreni"
[45] "aquifex_aeolicus"            "nanoarchaeum_equitans"
[47] "sulfolobus_islandicus"        "thermococcus_eurythermalis"
[49] "thermococcus_nautili"         "thermococcus_ornurineus"
[51] "thermofilum_adornatus"        "thermofilum_carboxyditrophus"
[53] "thermofilum_uzonense"         "thermogladus_cellulolyticus"

(todo <- unique(tdd[is.na(tdd$domain), "genre"]))

[1] "Anaplasma"           "Asticcacaulis"       "Borrelia"
[4] "Buchnera"           "Caulobacter"         "Chlamydophila"
[7] "Chlorobium"         "Chloroherpeton"     "Coxiella"
[10] "Ehrlichia"          "Hahella"             "Halanaeroarchaeum"
[13] "Halanaerobium"     "Halonotius"         "Haloquadratum"
[16] "Halotheca"         "Mannheimia"         "Mycobacterium"
[19] "Nitrospina"        "Nodularia"          "Orientia"
[22] "Propionibacterium" "Ralstonia"          "Rhodococcus"
[25] "Rickettsia"        "Ruegeria"           "Thermosynechococcus"
[28] "Tolypothrix"       "Tropheryma"         "Vibrio"
[31] "Wigglesworthia"   "Wolbachia"          "Cenarchaeum"
[34] "Halothermothrix"  "Aquifex"            "Nanoarchaeum"
[37] "Sulfolobus"       "Thermococcus"       "Thermofilum"
[40] "Thermogladus"

quoi <- c("organism", "domain")
for(g in todo){
  if(all(is.na(tdd[tdd$genre == g, "domain"]))) {
    print(paste("raté pour", g))
  } else {
    onegood <- tdd[tdd$genre == g & !is.na(tdd$domain), ][1, ]
    tdd[tdd$genre == g & is.na(tdd$domain), 66:74] <- onegood[66:74]
  }
}

[1] "raté pour Anaplasma"
[1] "raté pour Buchnera"
[1] "raté pour Chloroherpeton"
[1] "raté pour Coxiella"
[1] "raté pour Ehrlichia"
[1] "raté pour Hahella"
[1] "raté pour Halanaeroarchaeum"
[1] "raté pour Halonotius"
[1] "raté pour Halotheca"
[1] "raté pour Nitrospina"
[1] "raté pour Nodularia"
[1] "raté pour Orientia"
[1] "raté pour Thermosynechococcus"
[1] "raté pour Tolypothrix"
[1] "raté pour Tropheryma"
[1] "raté pour Wigglesworthia"
[1] "raté pour Wolbachia"
[1] "raté pour Cenarchaeum"
[1] "raté pour Halothermothrix"
[1] "raté pour Aquifex"
[1] "raté pour Nanoarchaeum"
[1] "raté pour Thermogladus"

tdd[is.na(tdd$domain), c(1, 66:67)]

      organism domain taxid
170   anaplasma_marginale <NA>   NA
171   anaplasma_phagocytophilum <NA>   NA
377     buchnera_aphidicola <NA>   NA
547   chloroherpeton_thalassium <NA>   NA
674     coxiella_burnetii <NA>   NA
797     ehrlichia_canis <NA>   NA
798     ehrlichia_chaffeensis <NA>   NA
800     ehrlichia_ruminantium <NA>   NA

```

```

985          hahella_chejuensis <NA> NA
987 halanaeroarchaeum_sulfurireducens <NA> NA
1008          halonotius_sp <NA> NA
1021          halothece_sp <NA> NA
1453          nitrospina_gracilis <NA> NA
1463          nodularia_spumigena <NA> NA
1498          orientia_tsutsugamushi <NA> NA
2167          thermosynechococcus_sp <NA> NA
2198          tolypothrix_sp <NA> NA
2217          tropheryma_whipplei <NA> NA
2256          wigglesworthia_glossinidia <NA> NA
2259          wolbachia_endosymbiont <NA> NA
517          cenarchaeum_symbiosum <NA> NA
1022          halothermothrix_orenii <NA> NA
175          aquifex_aeolicus <NA> NA
1411          nanoarchaeum_equitans <NA> NA
2155          thermogladius_cellulolyticus <NA> NA

```

```
unique(tdd[is.na(tdd$domain), "genre"])
```

```

[1] "Anaplasma"          "Buchnera"          "Chloroherpeton"
[4] "Coxiella"          "Ehrlichia"        "Hahella"
[7] "Halanaeroarchaeum" "Halonotius"       "Halothece"
[10] "Nitrospina"       "Nodularia"        "Orientia"
[13] "Thermosynechococcus" "Tolypothrix"     "Tropheryma"
[16] "Wigglesworthia"   "Wolbachia"        "Cenarchaeum"
[19] "Halothermothrix" "Aquifex"          "Nanoarchaeum"
[22] "Thermogladius"

```

JE complète à la main mais juste pour le domaine, il faudra se rappeler pour la suite que la taxonomie complète n'est pas forcément renseignée. TODO compléter la taxonomie à la main, c'est trop pénible d'avoir des NA.

Check that there is no more hard-coding for column selection but always a `which(colnames(tdd) == target)`

```

tdd[tdd$organism == "anaplasma_marginale", "domain"] <- "Bacteria"
# anaplasmataceae TID 942
tdd[!is.na(tdd$family) & tdd$family == 942, "organism"] # empty

```

```
character(0)
```

```

tdd[tdd$organism == "anaplasma_marginale", "family"] <- 942
# rickettsiales TID 766
tdd[!is.na(tdd$order) & tdd$order == 766, c("organism", "order")] # rickettsia_conorii

```

```

          organism order
1792 rickettsia_conorii 766
1794 rickettsia_felis 766
1798 rickettsia_monacensis 766
1799 rickettsia_montanensis 766
1803 rickettsia_prowazekii 766
1805 rickettsia_rickettsii 766
1808 rickettsia_typhi 766

```

```
target0 <- c("superkingdom", "phylum", "class", "order")
```

```
tdd[tdd$organism == "anaplasma_marginale", target0] <- tdd[tdd$organism == "rickettsia_conorii", target0]
```

```
tdd[tdd$organism == "anaplasma_phagocytophilum", "domain"] <- "Bacteria"
```

```
targetF <- c("superkingdom", "phylum", "class", "order", "family")
```

```
tdd[tdd$organism == "anaplasma_phagocytophilum", targetF] <- tdd[tdd$organism == "anaplasma_marginale", targetF]
```

```
tdd[tdd$organism == "buchnera_aphidicola", "domain"] <- "Bacteria"
```

```
# erwiniaceae TID 1903409
```

```
tdd[!is.na(tdd$family) & tdd$family == 1903409, "organism"] # erwinia_amylovora
```

```

[1] "erwinia_amylovora" "erwinia_billingiae" "erwinia_pyrifoliae"
[4] "erwinia_sp"        "erwinia_tasmaniensis" "erwinia_tracheiphila"
[7] "pantoea_agglomerans" "pantoea_ananatis" "pantoea_rwandensis"
[10] "pantoea_sp"        "pantoea_stewartii" "pantoea_vagans"

```

```
tdd[tdd$organism == "buchnera_aphidicola", targetF] <- tdd[tdd$organism == "erwinia_amylovora", targetF]
```

```
tdd[tdd$organism == "chloroherpeton_thalassium", "domain"] <- "Bacteria"
```

```
# chlorobiaceae TID 191412
```

```
tdd[!is.na(tdd$family) & tdd$family == 191412, "organism"] # chlorobaculum_parvum
```

```

[1] "chlorobaculum_parvum" "chlorobium_chlorochromatii"
[3] "chlorobium_limicola" "chlorobium_phaeovibrioides"
[5] "chlorobium_tepidum"

```

```
tdd[tdd$organism == "chloroherpeton_thalassium", targetF] <- tdd[tdd$organism == "chlorobaculum_parvum", targetF]
```

```
tdd[tdd$organism == "coxiella_burnetii", "domain"] <- "Bacteria"
```

```
# coxiellaceae TID 118968
```

```
tdd[!is.na(tdd$family) & tdd$family == 118968, "organism"] # empty
```

```

character(0)
tdd[tdd$organism == "coxiella_burnetii", "family"] <- 118968
# legionellales TID 118969
tdd[!is.na(tdd$order) & tdd$order == 118969, c("organism", "order")] # legionella_fallonii
      organism order
1174 legionella_fallonii 118969
1175 legionella_hackeliae 118969
1176 legionella_longbeachae 118969
1177 legionella_oakridgensis 118969
1178 legionella_pneumophila 118969
2097 tatlockia_micdadei 118969

tdd[tdd$organism == "coxiella_burnetii", target0] <- tdd[tdd$organism == "legionella_fallonii", target0]
tdd[tdd$organism == "ehrlichia_canis", "domain"] <- "Bacteria"
# anaplasmataceae TID 942
tdd[!is.na(tdd$family) & tdd$family == 942, "organism"] # legionella_fallonii

[1] "anaplasma_marginale"      "anaplasma_phagocytophilum"

tdd[tdd$organism == "ehrlichia_canis", targetF] <- tdd[tdd$organism == "legionella_fallonii", targetF]
tdd[tdd$organism == "ehrlichia_chaffeensis", "domain"] <- "Bacteria"
tdd[tdd$organism == "ehrlichia_chaffeensis", targetF] <- tdd[tdd$organism == "legionella_fallonii", targetF]
tdd[tdd$organism == "ehrlichia_ruminantium", "domain"] <- "Bacteria"
tdd[tdd$organism == "ehrlichia_ruminantium", targetF] <- tdd[tdd$organism == "legionella_fallonii", targetF]
tdd[tdd$organism == "hahella_chejuensis", "domain"] <- "Bacteria"
# hahellaceae TID 224379
tdd[!is.na(tdd$family) & tdd$family == 224379, "organism"] # empty

character(0)
tdd[tdd$organism == "hahella_chejuensis", "family"] <- 224379
# oceanospirillales TID 135619
tdd[!is.na(tdd$order) & tdd$order == 135619, "organism"] # alcanivorax_borkumensis

[1] "alcanivorax_borkumensis"      "alcanivorax_dieselolei"
[3] "alcanivorax_sp"              "bermanella_marisrubri"
[5] "chromohalobacter_salexigens" "halomonas_boliviensis"
[7] "halomonas_campaniensis"     "halomonas_elongata"
[9] "halomonas_sp"               "kangiella_koreensis"
[11] "marinomonas_mediterranea"   "neptuniibacter_caesariensis"
[13] "reinekea_blandensis"        "thalassolituus_oleivorans"
[15] "marinomonas_sp"             "oleispira_antarctica"

tdd[tdd$organism == "hahella_chejuensis", target0] <- tdd[tdd$organism == "alcanivorax_borkumensis", target0]
tdd[tdd$organism == "nitrospina_gracilis", "domain"] <- "Bacteria"
# nitrospinaceae TID 407032
tdd[!is.na(tdd$family) & tdd$family == 407032, "organism"] # empty

character(0)
tdd[tdd$organism == "nitrospina_gracilis", "family"] <- 407032
# nitrosinales TID 1293499
tdd[!is.na(tdd$order) & tdd$order == 1293499, "organism"] # empty

character(0)
tdd[tdd$organism == "nitrospina_gracilis", "order"] <- 1293499
# nitrospina TID 1293498
tdd[!is.na(tdd$class) & tdd$class == 1293498, "organism"] # empty

character(0)
tdd[tdd$organism == "nitrospina_gracilis", "class"] <- 1293498
# nitrospinae TID 1293497
tdd[!is.na(tdd$phylum) & tdd$phylum == 1293497, "organism"] # empty

character(0)
tdd[tdd$organism == "nitrospina_gracilis", "phylum"] <- 1293497
tdd[tdd$organism == "nitrospina_gracilis", "superkingdom"] <- 2
tdd[tdd$organism == "nodularia_spumigena", "domain"] <- "Bacteria"
# aphanizomenoaceae TID 1892259
tdd[!is.na(tdd$family) & tdd$family == 1892259, "organism"] # empty

character(0)
tdd[tdd$organism == "nodularia_spumigena", "family"] <- 1892259
# nostocales TID 1161
tdd[!is.na(tdd$order) & tdd$order == 1161, "organism"] # anabaena_cylindrica

[1] "anabaena_cylindrica" "anabaena_sp"      "anabaena_variabilis"
[4] "calothrix_sp"

```

```

tdd[tdd$organism == "nodularia_spumigena", target0] <- tdd[tdd$organism == "anabaena_cylindrica", target0]
tdd[tdd$organism == "orientia_tsutsugamushi", "domain"] <- "Bacteria"
# rickettsiaceae TID 775
tdd[tdd$organism == "orientia_tsutsugamushi", targetF] <- tdd[tdd$organism == "rickettsia_conorii", targetF]
tdd[tdd$organism == "tolypothrix_sp", "domain"] <- "Bacteria"
# tolypothrichaceae TID 119859
tdd[!is.na(tdd$family) & tdd$family == 119859, "organism"] # empty
character(0)
tdd[tdd$organism == "tolypothrix_sp", "family"] <- 119859
# nostocales TID 1161
tdd[tdd$organism == "tolypothrix_sp", target0] <- tdd[tdd$organism == "anabaena_cylindrica", target0]
tdd[tdd$organism == "tropheryma_whipplei", "domain"] <- "Bacteria"
# unclassified micrococcales TID 577468
tdd[!is.na(tdd$family) & tdd$family == 577468, "organism"] # empty
character(0)
tdd[tdd$organism == "tropheryma_whipplei", "family"] <- 577468
# micrococcales TID 85006
tdd[!is.na(tdd$order) & tdd$order == 85006, "organism"] # arthrobacter_sp
[1] "arthrobacter_sp" "beutenbergia_cavernae"
[3] "brachybacterium_faecium" "brachybacterium_muris"
[5] "brachybacterium_phenoliresistens" "cellulomonas_fimi"
[7] "cellulomonas_flavigena" "clavibacter_michiganensis"
[9] "curtobacterium_flaccumfaciens" "dermabacter_hominis"
[11] "dermacoccus_nishinomiyaensis" "glutamicibacter_arilaitensis"
[13] "intrasporangium_calvum" "isopterocola_variabilis"
[15] "janibacter_sp" "jonesia_denitrificans"
[17] "kocuria_rhizophila" "kocuria_sp"
[19] "kytrococcus_sedentarius" "leifsonia_aquatica"
[21] "leifsonia_xyli" "leucobacter_sp"
[23] "microbacterium_sp" "microbacterium_testaceum"
[25] "micrococcus_luteus" "paenarthrobacter_aurescens"
[27] "pseudarthrobacter_chlorophenolicus" "rothia_aeria"
[29] "rothia_dentocariosa" "rothia_mucilaginoso"
[31] "sanguibacter_keddiei" "tetrasphaera_elongata"
[33] "xylanimonas_cellulosilytica" "renibacterium_salmoninarum"
tdd[tdd$organism == "tropheryma_whipplei", target0] <- tdd[tdd$organism == "arthrobacter_sp", target0]
tdd[tdd$organism == "wigglesworthia_glossinidia", "domain"] <- "Bacteria"
# erwiniaaceae TID 1903409
tdd[!is.na(tdd$family) & tdd$family == 1903409, "organism"] # erwinia_amylovora
[1] "buchnera_aphidicola" "erwinia_amylovora" "erwinia_billingiae"
[4] "erwinia_pyriifoliae" "erwinia_sp" "erwinia_tasmaniensis"
[7] "erwinia_tracheiphila" "pantoea_agglomerans" "pantoea_ananatis"
[10] "pantoea_rwandensis" "pantoea_sp" "pantoea_stewartii"
[13] "pantoea_vagans"
tdd[tdd$organism == "wigglesworthia_glossinidia", targetF] <- tdd[tdd$organism == "erwinia_amylovora", targetF]
tdd[tdd$organism == "wolbachia_endosymbiont", "domain"] <- "Bacteria"
# anaplasmataceae TID 942
tdd[!is.na(tdd$family) & tdd$family == 942, "organism"] # anaplasma_marginale
[1] "anaplasma_marginale" "anaplasma_phagocytophilum"
tdd[tdd$organism == "wolbachia_endosymbiont", targetF] <- tdd[tdd$organism == "anaplasma_marginale", targetF]
tdd[tdd$organism == "cenarchaeum_symbiosum", "domain"] <- "Archaea"
# cenarchaeaceae TID 205957
tdd[!is.na(tdd$family) & tdd$family == 205957, "organism"] #empty
character(0)
tdd[tdd$organism == "cenarchaeum_symbiosum", "family"] <- 205957
# cernarchales TID 205948
tdd[!is.na(tdd$order) & tdd$order == 205948, "organism"] #empty
character(0)
tdd[tdd$organism == "cenarchaeum_symbiosum", "order"] <- 205948
# pas de class
tdd[tdd$organism == "cenarchaeum_symbiosum", "class"] <- 0
tdd[tdd$organism == "cenarchaeum_symbiosum", "phylum"] <- 651137
tdd[tdd$organism == "cenarchaeum_symbiosum", "superkingdom"] <- 2157
tdd[tdd$organism == "aquifex_aeolicus", "domain"] <- "Bacteria"
# aquificaceae TID 64898
tdd[!is.na(tdd$family) & tdd$family == 64898, "organism"] # thermocrinis_albus
[1] "hydrogenobacter_thermophilus" "thermocrinis_albus"
[3] "thermocrinis_ruber"

```

```

tdd[tdd$organism == "aquifex_aeolicus", targetF] <- tdd[tdd$organism == "thermocrinis_albus", targetF]
tdd[tdd$organism == "nanoarchaeum_equitans", "domain"] <- "Archaea"
# nanoarchaeaceae TID 1890941
tdd[!is.na(tdd$family) & tdd$family == 1890941, "organism"] # empty

character(0)

tdd[tdd$organism == "nanoarchaeum_equitans", "family"] <- 1890941
# nanoarchaeales TID 1890940
tdd[!is.na(tdd$order) & tdd$order == 1890940, "organism"] # empty

character(0)

tdd[tdd$organism == "nanoarchaeum_equitans", "order"] <- 1890940
# pas de class
tdd[tdd$organism == "nanoarchaeum_equitans", "class"] <- 0
tdd[tdd$organism == "nanoarchaeum_equitans", "phylum"] <- 192989
tdd[tdd$organism == "nanoarchaeum_equitans", "superkingdom"] <- 2157
tdd[is.na(tdd$domain), 1]

[1] "halanaeroarchaeum_sulfurireducens" "halonotius_sp"
[3] "halothece_sp" "thermosynechococcus_sp"
[5] "halothermothrix_oreonii" "thermogladius_cellulolyticus"

tdd$domain <- factor(tdd$domain)

tdd[is.na(tdd$family), c("organism", targetF)]

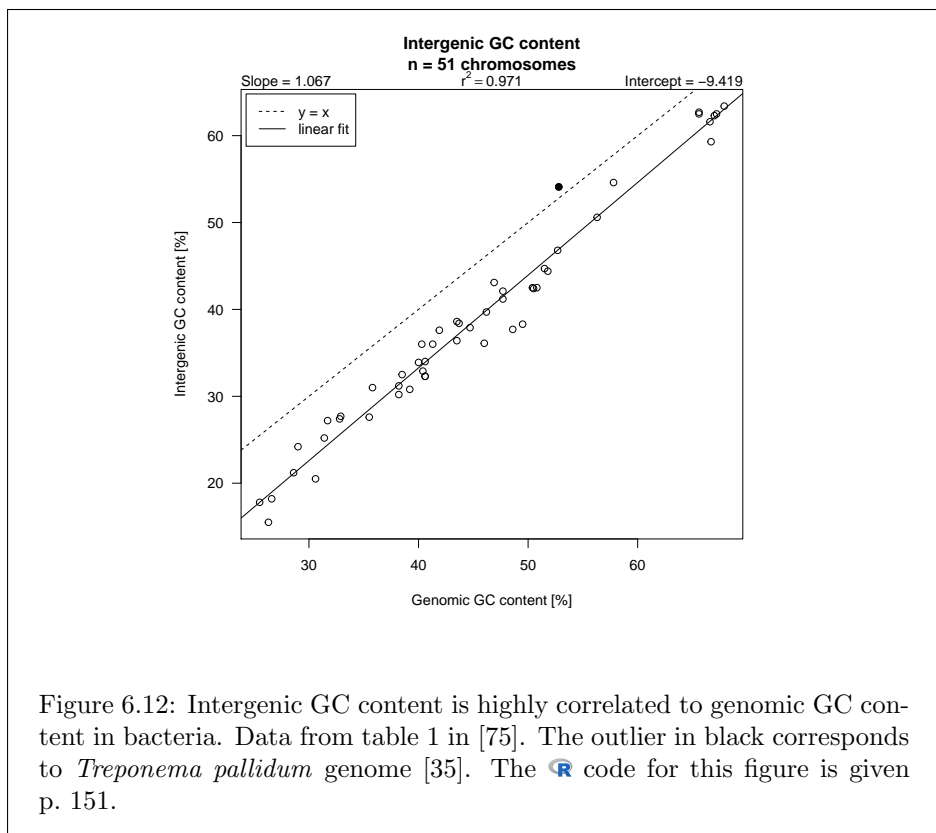
```

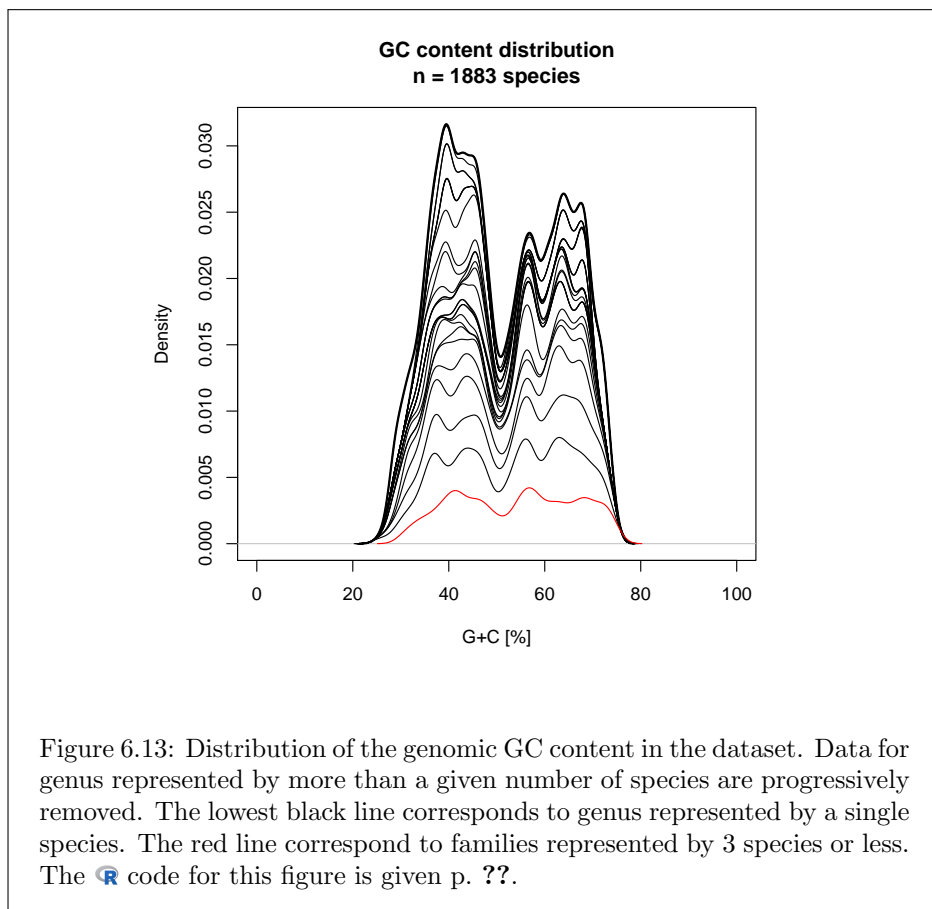
	organism	superkingdom	phylum	class	order	family
250	bacteroides_pectinophilus	2	1239	186801	186802	NA
536	chloracidobacterium_thermophilum	2	57723	1562566	NA	NA
566	clostridium_acidurici	2	1239	186801	186802	NA
596	clostridium_orbiscindens	2	1239	186801	186802	NA
610	clostridium_ultunense	2	1239	1737404	NA	NA
701	defluviiitoga_tunisiensis	2	200918	188708	1643947	NA
873	exiguobacterium_antarcticum	2	1239	91061	1385	NA
874	exiguobacterium_sibiricum	2	1239	91061	1385	NA
875	exiguobacterium_sp	2	1239	91061	1385	NA
902	flavonifractor_plautii	2	1239	186801	186802	NA
933	gemella_bergeriae	2	1239	91061	1385	NA
934	gemella_haemolysans	2	1239	91061	1385	NA
1191	leptothrix_cholodnii	2	1224	28216	80840	NA
1324	methylibium_petroleiphilum	2	1224	28216	80840	NA
1333	methyloceanibacter_caenitepidi	2	1224	28211	356	NA
1592	petrotoga_mobilis	2	200918	188708	1643947	NA
1615	pleisiomonas_shigelloides	2	1224	1236	91347	NA
1832	rubrivivax_gelatinosus	2	1224	28216	80840	NA
2080	sulfurovum_lithotrophicum	2	1224	29547	NA	NA
2129	thermobaculum_terrenum	2	NA	NA	NA	NA
2131	thermobispora_bispora	2	201174	1760	NA	NA
2188	thiolapillus_brandeum	2	1224	1236	NA	NA
2190	thiomonas_arsenitoxydans	2	1224	28216	80840	NA
2191	thiomonas_intermedia	2	1224	28216	80840	NA
2258	wohlfahrtiimonas_chitiniclastica	2	1224	1236	NA	NA
38	aciduliprofundum_boonei	2157	28890	NA	NA	NA
1237	marinitoga_piezophila	2	200918	188708	1643947	NA
1438	nitratifractor_salsuginis	2	1224	29547	213849	NA

IL me reste quand même 28 bactérie dont la taxonomie est partielle. Je décide de les garder quand même, tant pis pour les NA.

6.4.2 GC content computation

THE GC content in coding sequences is used here as a proxy for the genomic GC content. The approximation in bacteria is not too bad because most of bacterial genome consists of coding sequences, typically ~90% [83]. Figure 6.12 page 136 shows that intergenic GC content is highly correlated ($r^2 \approx 0.97$) to the genomic GC content but systematically lower by 10%. Then, the proxy used here overestimate the actual genomic GC content by a ~1% unit, for instance a 55% value should be in fact 54%.





```
codons <- colnames(tdd[,2:65])
ngc <- function(x){
  sum(s2c(x) %in% c("c", "g"))/3
}
alpha <- sapply(codons, ngc)
freq <- as.matrix(tdd[,2:65]/rowSums(tdd[,2:65]))
tdd$tdgc <- (100*freq %*% alpha)[ , 1]
```

THE distribution of the GC content in the dataset is given in figure 6.13 page 137. The distribution of the GC content looks like a mixture of two normal distributions, suggesting an underlying qualitative variable. Potential candidates to investigate include:

- Aerobiosis [90].
- The pol III α subunit [160].
- Environment [114].
- To be continued...

6.4.3 Computing aminoacid frequencies

TO compute aminoacid frequencies we just define a factor `facaa` for the codons giving the encoded aminoacid. The function `tapply()` is used then to compute the total number of codons by aminoacid. Figure 6.14 page 139 gives the distribution of aminoacid frequencies in the dataset.

```
codons <- colnames(tdd[ , 2:65])
codons
[1] "aaa" "aac" "aag" "aat" "aca" "acc" "acg" "act" "aga" "agc" "agg" "agt" "ata"
[14] "atc" "atg" "att" "caa" "cac" "cag" "cat" "cca" "ccc" "ccg" "cct" "cga" "cgc"
[27] "cgg" "cgt" "cta" "ctc" "ctg" "ctt" "gaa" "gac" "gag" "gat" "gca" "gcc" "gcg"
[40] "gct" "gga" "ggc" "ggg" "ggg" "ggt" "gta" "gtc" "gtg" "gtt" "taa" "tac" "tag" "tat"
[53] "tca" "tcc" "tcg" "tct" "tga" "tgc" "tgg" "tgt" "tta" "ttc" "ttg" "ttt"

facaa <- factor(sapply(codons, function(x) aaa(translate(s2c(x)))))
facaa
aaa aac aag aat aca acc acg act aga agc agg agt ata atc atg att caa cac cag cat cca
Lys Asn Lys Asn Thr Thr Thr Arg Ser Arg Ser Ile Ile Met Ile Gln His Gln His Pro
ccc ccg cct cga cgc cgg cgt cta ctc ctg ctt gaa gac gag gat gca gcc gcg gct gga gcc
Pro Pro Pro Arg Arg Arg Arg Leu Leu Leu Leu Glu Asp Glu Asp Ala Ala Ala Ala Gly Gly
ggg ggt gta gtc gtg gtt taa tac tag tat tca tcc tcg tct tga tgc tgg tgt tta ttc ttg
Gly Gly Val Val Val Val Stp Tyr Stp Tyr Ser Ser Ser Ser Stp Cys Trp Cys Leu Phe Leu
ttt
Phe
21 Levels: Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser ... Val

aa <- t(apply(tdd[ , 2:65], 1, function(x) tapply(x, facaa, sum)))
aa <- 100*aa/rowSums(aa) # relative frequencies in percent
tdd <- cbind(tdd, aa)
```

Discuss outliers somewhere

6.4.4 Computing isoelectric points

Check pI typography in preamble

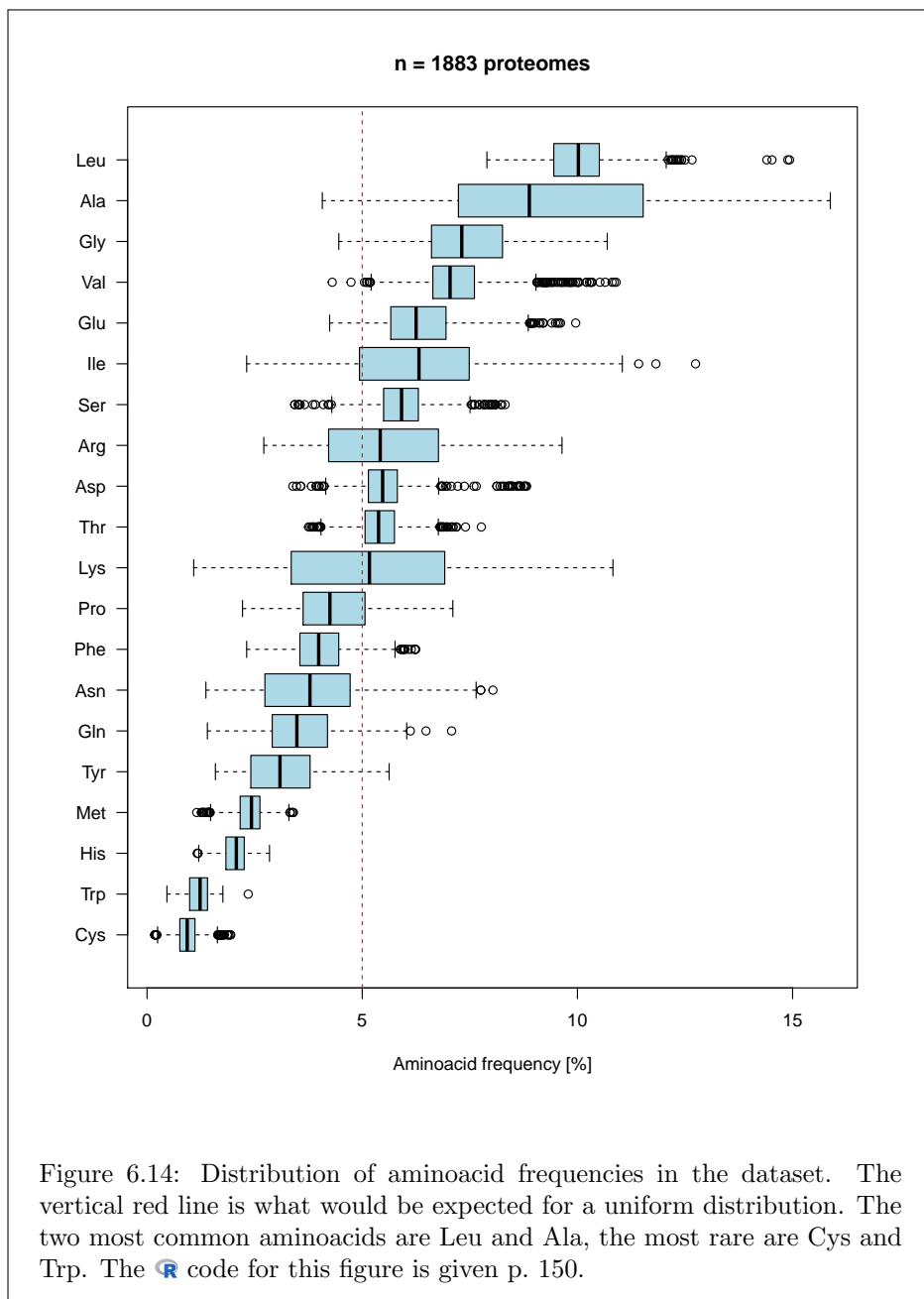
BY definition the isoelectric point, pI , is the pH value for which positive charges are cancelled by negative charges: at $pH = pI$ the sum of charges is zero. In post-transcriptionally unprocessed proteins there are four ionisable groups that can have a positive charge: the lysin (K), arginin (R) and histidine (H) residuals and the N-terminal $-NH_2$. There are five ionisable groups that can have a negative charge: tyrosin (Y), cystein (C), aspartate (D) and glutamate (E) residuals and the C-terminal $-COOH$. In the following we neglect the N- and C-terminal groups, but anyway we also neglect all post-transcriptional modifications.

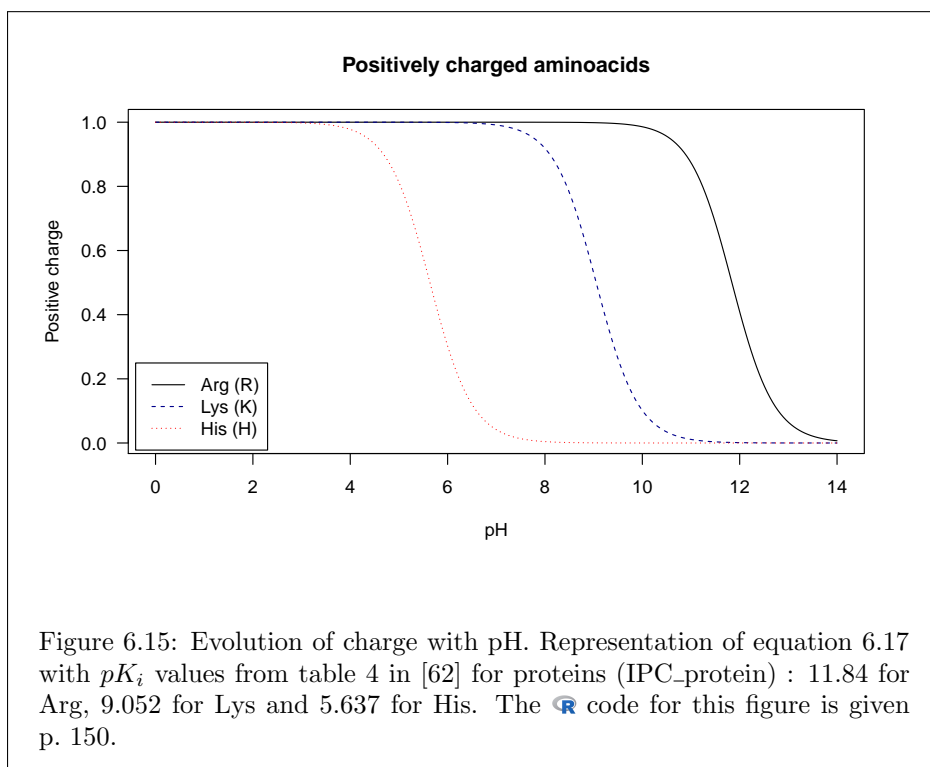
LET $f^+(pH)$ be the sum of all positive charges and $f^-(pH)$ the sum of all negative charges for a given pH. Let $I^+ = \{K, R, H\}$ the set of positively charged residuals and $I^- = \{Y, C, D, E\}$ the set of negatively charged residuals. Therefore, the positive charge, $f^+(pH)$ of a protein is given by:

$$f^+(pH) = \sum_{i \in I^+} n_i f_i^+(pH) \quad (6.14)$$

where $f_i^+(pH)$ is the positive charge for the aminoacid of kind i and n_i the total number of aminoacids of this type in the protein. In a similar way, the negative charge, $f^-(pH)$ of a protein is given by:

$$f^-(pH) = \sum_{i \in I^-} n_i f_i^-(pH) \quad (6.15)$$





where $f_i^-(\text{pH})$ is the negative charge for the aminoacid of kind i and n_i the total number of aminoacids of this type in the protein. From the definition of the pI , the equation we want to solve is:

$$f^+(\text{pH}) + f^-(\text{pH}) = 0 \quad (6.16)$$

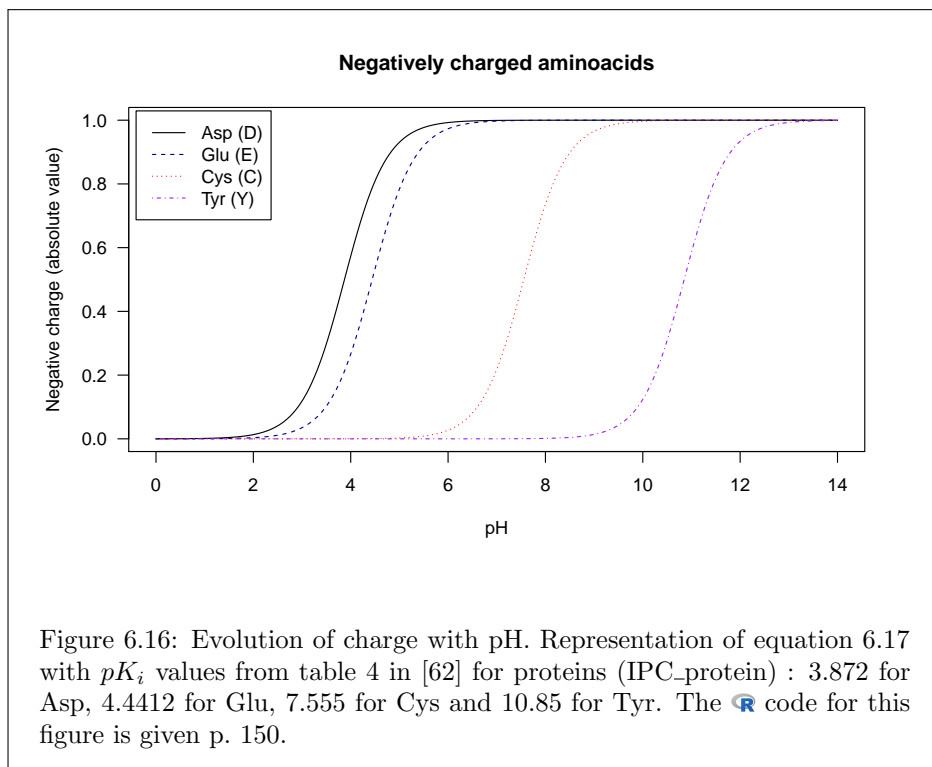
IN order to solve equation 6.16 we need the analytical expression for $f_i^+(\text{pH})$ and $f_i^-(\text{pH})$. These are parametric functions, with a single parameter pK_i for aminoacid i , given by the so-called HENDERSON-HASSELBACH *approximation* [48, 47, 105, 26]. For positively charged aminoacids we have:

$$f_i^+(\text{pH}) = \frac{1}{1 + 10^{\text{pH} - pK_i}} \quad (6.17)$$

and the corresponding curves are given in figure 6.15 page 140. They are all monotonously decreasing from 1 to 0 sigmoids with an inflection point at $\text{pH} = pK_i$. The inflection point is a symmetric point such that $f_i^+(pK_i) = \frac{1}{2}$. At neutral pH, only Arg and Lys are positively charged. For negatively charged aminoacids we have:

$$f_i^-(\text{pH}) = \frac{10^{\text{pH} - pK_i}}{1 + 10^{\text{pH} - pK_i}} \quad (6.18)$$

and the corresponding curves are given in figure 6.16 page 141. They are all monotonously increasing from 0 to 1 (absolute value) sigmoids with an inflection point at $\text{pH} = pK_i$. The inflection point is a symmetric point such that $f_i^-(pK_i) = \frac{1}{2}$. At neutral pH, only Asp and Glu are negatively charged.

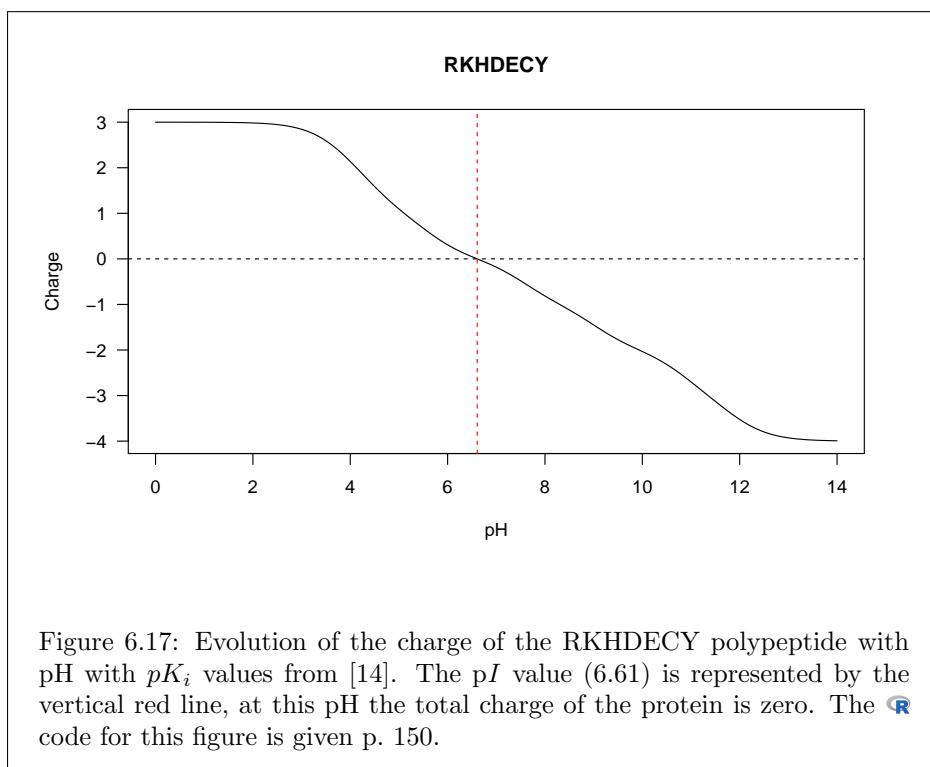


BEFORE solving equation 6.16 in general, let's have a look of a simpler case. We consider a small protein with one representative of all ionisable aminoacids, that is the polypeptide RKHDECY. Figure 6.17 page 142 shows that the charge monotonously decrease from +3 to -4 passing 0 at $pH = pI$. In the general case, the evolution of charge with pH is described by a linear combination of the seven underlying functions. There is no analytical solution to equation 6.16, but the standard `R` `uniroot()` is at hand for this. We have already computed aminoacids frequencies in the previous section, we can use them to get the pI :

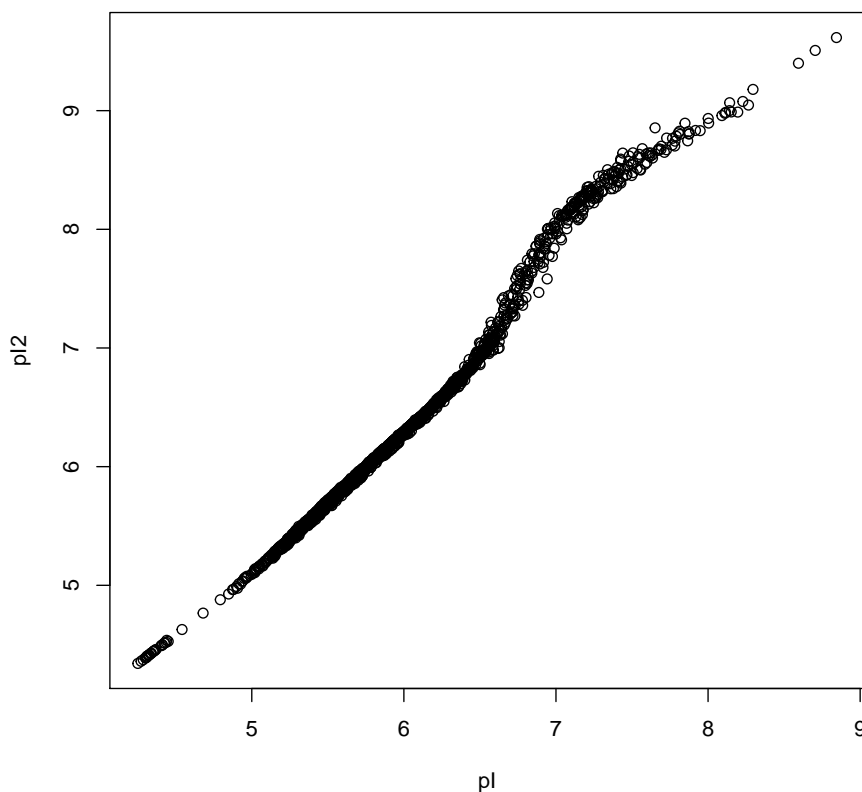
Find an mnemonic anagram. HERDCYK ?

```
getpI <- function(faa, pK){
  computeCharge <- function(pH, faa, pK){
    pos <- faa["R"]*computePositiveCharge("R", pH, pK) +
      faa["K"]*computePositiveCharge("K", pH, pK) +
      faa["H"]*computePositiveCharge("H", pH, pK)
    neg <- faa["D"]*computeNegativeCharge("D", pH, pK) +
      faa["E"]*computeNegativeCharge("E", pH, pK) +
      faa["C"]*computeNegativeCharge("C", pH, pK) +
      faa["Y"]*computeNegativeCharge("Y", pH, pK)
    return(pos - neg)
  }
  return(uniroot(computeCharge, c(0,14), faa = faa, pK = pK)$root)
}
faautil <- tdd[, colnames(tdd) %in% aaa(s2c("RKHDECY"))]
colnames(faautil) <- a(colnames(faautil))
tdd$pI <- apply(faautil, 1, getpI, pK = pK)

pK2 <- SEQINR.UTIL$pk[,3]
names(pK2) <- rownames(SEQINR.UTIL$pk)
pK2 <- pK2[names(pK2) %in% s2c("CDEHKRY")]
tdd$pI2 <- apply(faautil, 1, getpI, pK = pK2)
with(tdd, plot(pI, pI2))
with(tdd, all.equal(pI, pI2))
```



[1] "Mean relative difference: 0.06016203"



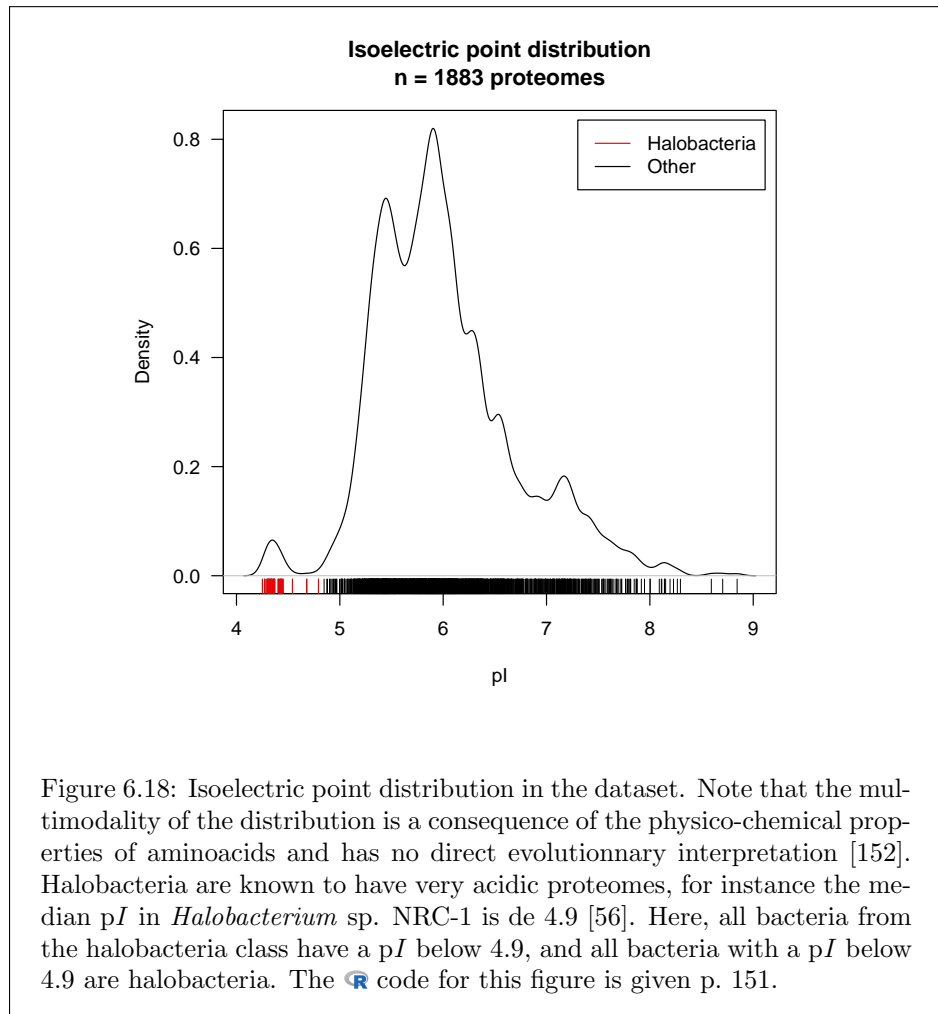
Try empirical models to have a good starting guess for pI

A kick-and-dirty ugly code

MY first attempt was to use directly the function `computePI()` from package `seqinr` [23] based on pK values from [14]. The expected input for this function is protein sequence as a vector of single chars in upper case. The kick-and-dirty approach is to generate from aminoacid frequencies a huge protein reflecting the proteome composition. This is very inefficient because we generate a huge object where we just need amino-acid relative frequencies. The execution time was in hours on my laptop.

```
# Compute aa absolute frequencies
facaac <- factor(sapply(codons, function(x) aaa(translate(s2c(x)))))
aaabs <- t(apply(tdd[,2:65], 1, function(x) tapply(x, facaac, sum)))
aaabs <- aaabs[, -which(colnames(aaabs) == "Stp")]
getPI <- function(x){
  prot <- rep(a(colnames(aaabs)), x)
  return(computePI(prot))
}
PIs <- apply(aaabs, 1, getPI)
save(PIs, file = "local/PIs.Rda")
```

Le point isoélectrique des protéomes a été calculé avec la fonction `La`. La multimodalité de la distribution des points isoélectriques est une conséquence



des propriétés physico-chimiques des acides-aminés et n'a pas d'interprétation d'un point de vue évolutif [152]. Les halobactéries sont connues pour avoir un protéome très acide, par exemple le pI médian chez *Halobacterium* sp. NRC-1 est de 4.9 [56]. Ici, toutes les bactéries de la classe des halobacteria ont un pI inférieur à 4.9.

Make forward biblio for proteome composition in Halobacteria

6.4.5 Backup

```
tdd$lineage_text <- as.character(tdd$lineage_text)
tdd <- tdd[order(tdd$organism), ]
save(tdd, file = "local/tdd.Rda")
```


Chapter 7

Conclusion

Chapter 8

Annexes

8.1 Code for figures



Code for figure 6.1 page 97.

```
models <- c("Square-root", "CTMI", "Blanchard", "MRM", "Hinshelwood",
           "DEB", "MMRM", "Proteome", "Heat-capacity", "Eppley-Norberg")
citations <- c(710, 267, 111, 107, 507, 000, 20, 126, 42, 146)
citations.2019.05.24 <- c(859, 333, 127, 129, 013, 777, 31, 179, 85, 187)
# Problem Hinshelwood, DEB
pbyear <- c(1983, 1993, 1996, 1946, 1946, 2010, 2014, 2011, 2014, 2004)
mcut <- as.data.frame(list(models = models, citations = citations, pbyear = pbyear,
                          citations.2019.05.24 = citations.2019.05.24))
mcut <- mcut[which(mcut$models %in% c("Hinshelwood", "DEB")), ]
mcut <- mcut[order(mcut$citations), ]
par(mar = c(5,8,0,1)+0.1)
bp <- barplot(mcut$citations, names = mcut$models, horiz = TRUE, las = 1, col = "lightblue",
             xlab = "Number of citations", xlim = c(0, 1000))
barplot(mcut$citations.2019.05.24, add = TRUE, horiz = TRUE, col = rgb(0.9,0.5,0.5,0.1))
text(mcut$citations.2019.05.24, bp[,1], mcut$pbyear, pos = 4)
```



Code for figure 6.2 page 107.

```
col <- rep("black", nrow(tocuT2))
tocuT2$delta <- with(tocuT2, abs(toptsp - temperature))
col[tocuT2$delta > 5] <- "orange"
col[tocuT2$delta > 10] <- "red"
n <- with(tocuT2, sum(!is.na(delta)))
main <- paste("Optimal growth temperature comparison\nn =", n)
with(tocuT2, plot(toptsp, temperature, asp = 1, xlab = "Topt [°C]",
                ylab = "Topt [°C]", las = 1, pch = 19, cex = 1, col = col,
                main = main),
     xlim = c(0, 100), ylim = c(0, 100))
abline(c(0, 1))
abline(c(5, 1), lty = 2); abline(c(-5, 1), lty = 2)
abline(c(10, 1), lty = 3); abline(c(-10, 1), lty = 3)
```



Code for figure 6.3 page 109.

```
y.scale <- 0.4/1377
temp <- c(65, 72.5, 79, 83, 87)
temp2 <- temp[-length(temp)] # Last temperature missing for strain P1
MT4 <- y.scale*c(230, 432, 778, 773, 493)
B12 <- y.scale*c(422, 630, 816, 569, 38)
P2 <- y.scale*c(385, 652, 903, 744, 387)
P1 <- y.scale*c(195, 541, 716, 596)
plot(temp, MT4, pch = 19, ylim = c(0, 0.3), las = 1, xlim = c(65, 88),
     xlab = "Temperature [°C]", ylab = "Mass - Doubling per hour",
     main = "Growth rates as function of temperature")
points(temp, B12, pch = 0, col = "darkblue")
points(temp, P2, pch = 2, col = "red3")
points(temp2, P1, pch = 17, col = "green4")
nlm.MT4 <- nlm.CTMI.auto(cbind(temp, MT4))
xx <- seq(from = 50, to = 90, length.out = 100)
y.MT4 <- sapply(xx, function(x) CTMI(x, nlm.MT4$estimate))
lines(xx, y.MT4)
```

```

nlm.B12 <- nlm.CTMI(c(55, 82, 87, 0.25), cbind(temp, B12))
y.B12 <- sapply(xx, function(x) CTMI(x, nlm.B12$estimate))
lines(xx, y.B12, lty = 2, col = "darkblue")
nlm.P2 <- nlm.CTMI(c(50, 80, 90, 0.25), cbind(temp, P2))
y.P2 <- sapply(xx, function(x) CTMI(x, nlm.P2$estimate))
lines(xx, y.P2, lty = 3, col = "red3")
nlm.P1 <- nlm.CTMI.auto(cbind(temp2, P1))
y.P1 <- sapply(xx, function(x) CTMI(x, nlm.P1$estimate))
lines(xx, y.P1, lty = 4, col = "green4")
legend("topleft", inset = 0.02, pch = c(19, 0, 2, 17), legend = c("MT4", "B12", "P2", "P1"),
      lty = c(1, 2, 3, 4), col = c("black", "darkblue", "red3", "green4"))

```



Code for figure 6.4 page 110.

```

x.MT4 <- c(60, 30*c(83, 93, 110, 130, 137, 145, 154)/159 + 60, 92)
y.MT4 <- c(0, log(2)*10^(-c(43, 31, 25, 0, -10, -15, 13)/97 - 1), 0)
x.MT3 <- c(45, 30*c(0, 29, 56, 83, 94)/159 + 60, 83)
y.MT3 <- c(0, log(2)*10^(-c(52, 28, 8, -14, 46)/97 - 1), 0)
plot(x.MT4, y.MT4, pch = 19, las = 1,
     xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
     main = "Growth rates as function of temperature", ylim = c(0, 0.12),
     xlim = c(40, 100))
points(x.MT3, y.MT3, pch = 1)
nlm.MT4 <- nlm.CTMI.auto(cbind(x.MT4, y.MT4))
xx <- seq(from = 50, to = 100, length.out = 500)
yy.MT4 <- sapply(xx, function(x) CTMI(x, nlm.MT4$estimate))
lines(xx, yy.MT4)
nlm.MT3 <- nlm.CTMI.auto(cbind(x.MT3, y.MT3))
xx <- seq(from = 40, to = 90, length.out = 500)
yy.MT3 <- sapply(xx, function(x) CTMI(x, nlm.MT3$estimate))
lines(xx, yy.MT3, lty = 2, col = "darkblue")
abline(h = 0)
legend("topleft", inset = 0.02, pch = c(19, 1), legend = c("MT4", "MT3"),
      lty = c(1, 2), col = c("black", "darkblue"))

```



Code for figure 6.5 page 112.

```

x.ca <- 30 + 40*c(29, 37, 55, 81, 108, 131, 156, 172, 178)/200
y.ca <- 10^(-0.699 - c(93, 67, 50, 42, 29, 20, 9, 39, 62)/81)
plot(x.ca, y.ca, pch = 4, las = 1,
     xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
     main = "Growth rates as function of temperature", ylim = c(0, 0.17),
     xlim = c(20, 70))
nlm.ca <- nlm.CTMI.auto(cbind(x.ca, y.ca))
xx <- seq(from = 20, to = 70, length.out = 500)
yy.ca <- sapply(xx, function(x) CTMI(x, nlm.ca$estimate))
lines(xx, yy.ca)
abline(h = 0)

```



Code for figure 6.6 page 114.

```

x.pa <- c(75, 80, 85, 90, 95, 100, 105, 109)
y.pa <- c(0.23, 0.32, 0.43, 0.63, 0.87, 0.90, 0.74, 0)
plot(x.pa, y.pa, pch = 19, las = 1,
     xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
     main = "Growth rates as function of temperature", ylim = c(0, 1),
     xlim = c(60, 110))
nlm.pa <- nlm.CTMI.auto(cbind(x.pa, y.pa))
xx <- seq(from = 60, to = 110, length.out = 500)
yy.pa <- sapply(xx, function(x) CTMI(x, nlm.pa$estimate))
lines(xx, yy.pa)

```



Code for figure 6.7 page 116.

```

x.ba <- c(17, 18, 20, 22, 25, 30, 35, 37, 40, 42, 43, 44, 25, 30, 35, 40)
y.ba <- c(11, 49, 64, 61, 153, 221, 301, 496, 488, 324, 102, 45, 164, 275, 362, 460)/1000
pch <- rep(c(18, 0), c(12, 4))
plot(x.ba, y.ba, pch = pch, las = 1,
     xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
     main = "Growth rates as function of temperature", xlim = c(10, 50),
     ylim = c(0, max(y.ba)))
nlm.ba <- nlm.CTMI.auto(cbind(x.ba, y.ba))
xx <- seq(from = 10, to = 50, length.out = 500)
yy.ba <- sapply(xx, function(x) CTMI(x, nlm.ba$estimate))
lines(xx, yy.ba)
legend("topleft", inset = 0.02, legend = c("Sterne", "Ames K0610"), pch = c(18, 0))

```



Code for figure 6.8 page 117.

```

x <- c(30, 35, 40, 45, 50, 55, 57, 60)
y.OH5 <- c(0.28, 1.34, 1.84, 2.24, 2.20, 1.43, 0.66, 0)
y.OH9 <- c(0.27, 1.05, 1.92, 2.51, 2.20, 2.05, 1.24, 0)
y.OH10 <- c(0.67, 1.55, 2.10, 2.40, 2.26, 1.46, 1.17, 0)
y.OH14 <- c(0.75, 1.43, 1.94, 2.42, 2.05, 1.65, 1.40, 0)
y.OH18 <- c(0.76, 1.81, 2.19, 2.85, 1.93, 1.49, 2.00, 0)
x.II <- c(35, 40, 45, 55, 57, 61, 63)
y.OH4 <- c(0, 0.42, 0.83, 1.11, 1.73, 0.99, 0)
y.OH20 <- c(0, 0.67, 0.94, 1.09, 1.57, 0.81, 0)
x.III <- c(40, 45, 55, 57, 61, 63, 65)
y.OH2 <- c(0, 0.74, 1.36, 1.60, 1.36, 0.93, 0)
x.IV <- c(50, 55, 61, 65, 70, 73)
y.OH28 <- c(0, 0.26, 0.63, 0.64, 0.69, 0)
y.OH29 <- c(0, 0.45, 0.55, 0.74, 0.54, 0)
y.OH30 <- c(0, 0.36, 0.41, 0.77, 0.69, 0)
plot(x, y.OH5, las = 1, pch = 1,
      xlab = "Temperature [°C]", ylab = "Specific growth rate [1/day]",
      main = "Growth rates as function of temperature", xlim = c(20, 80),
      ylim = c(0, 3))
points(x, y.OH9, pch = 2)
points(x, y.OH10, pch = 3)
points(x, y.OH14, pch = 4)
points(x, y.OH18, pch = 5)
points(x.II, y.OH4, pch = 6, col = "darkblue")
points(x.II, y.OH20, pch = 7, col = "darkblue")
points(x.III, y.OH2, pch = 8, col = "red")
points(x.IV, y.OH28, pch = 9, col = "green4")
points(x.IV, y.OH29, pch = 10, col = "green4")
points(x.IV, y.OH30, pch = 11, col = "green4")
nlm.I <- nlm.CTMI(c(20, 45, 60, 2.5), cbind(rep(x, 5), c(y.OH5, y.OH9, y.OH10, y.OH14, y.OH18)))
nlm.II <- nlm.CTMI.auto(cbind(rep(x.II, 2), c(y.OH4, y.OH20)))
nlm.III <- nlm.CTMI.auto(cbind(x.III, y.OH2))
nlm.IV <- nlm.CTMI.auto(cbind(rep(x.IV, 3), c(y.OH28, y.OH29, y.OH30)))
xx <- seq(from = 20, to = 80, length.out = 500)
yy.I <- sapply(xx, function(x) CTMI(x, nlm.I$estimate))
yy.II <- sapply(xx, function(x) CTMI(x, nlm.II$estimate))
yy.III <- sapply(xx, function(x) CTMI(x, nlm.III$estimate))
yy.IV <- sapply(xx, function(x) CTMI(x, nlm.IV$estimate))
lines(xx, yy.I)
lines(xx, yy.II, col = "darkblue")
lines(xx, yy.III, col = "red")
lines(xx, yy.IV, col = "green4")
mycols <- c("black", "darkblue", "red", "green4")
legend("topleft", inset = 0.02,
      legend = c("OH5", "OH9", "OH10", "OH14", "OH18", "OH4", "OH20", "OH2",
                 "OH28", "OH29", "OH30"), pch = 1:11, col = rep(mycols, c(5,2,1,3)))
legend("topright", inset = 0.02,
      legend = c("Group I", "Group II", "Group III", "Group IV"),
      lty = 1, col = mycols)

```



Code for figure 6.9 page 118.

```

x.hv <- seq(from = 23, to = 49, by = 2) # consistent with text and figure
pixs <- c(282, 341, 389, 431, 495, 521, 545, 589, 613, 641, 672, 691, 662, 566)
# Pixels above ln(k) = -2 on my screenshot
lnk <- -2 + 2*pixs/869 # I have 2 log units corresponding to 869 pixels
par(mfrow = c(1, 2), xaxs = "i")
# Reproduce figure iE
plot(-x.hv, lnk, ylim = c(-2, 0), xlim = c(-63, -19), las = 1, pch = 15,
      main = "Hfr: volcanii (DS70)", xlab = "Temperature (°C)",
      ylab = "log k", xaxt = "n", xaxs = "i", yaxs = "i", bty = "n")
tlab <- seq(from = 19, to = 59, by = 10)
par(xaxs = "i")
axis(1, at = -tlab, labels = tlab, xaxs = "i")
axis(1, at = -seq(from = 19, to = 63, by = 2), labels = NA, xaxs = "i", tcl = -0.25)
text(-49, lnk[14], "49", pos = 2)
text(-45, lnk[12], "45", pos = 3)
text(-31, lnk[5], "31", pos = 3)
AP <- 5:12 # Arrhenius portion
abline(lm(rev(lnk[AP])~rev(-x.hv[AP])), lty = "5313")
# Second figure with CTMI model
y.hv <- exp(lnk)
plot(x.hv, y.hv, las = 1, pch = 1,
      xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
      main = "Growth rate as function of temperature", xlim = c(0, 60),
      ylim = c(0, max(y.hv)))
nlm.hv <- nlm.CTMI.auto(cbind(x.hv, y.hv))
xx <- seq(from = 0, to = 60, length.out = 500)
yy.hv <- sapply(xx, function(x) CTMI(x, nlm.hv$estimate))
lines(xx, yy.hv)

```



Code for figure 6.10 page 119.

```

x.tn <- 40 + 60*c(11, 33, 54, 73, 80, 88, 104, 120)/129
y.tn <- c(0, 18, 79, 137, 145, 128, 60, 0)/144
plot(x.tn, y.tn, las = 1, pch = 19,
      xlab = "Temperature [°C]", ylab = "Specific growth rate [1/h]",
      main = "Growth rate as function of temperature", xlim = c(40, 100),
      ylim = c(0, max(y.tn)))
nlm.tn <- nlm.CTMI.auto(cbind(x.tn, y.tn))
xx <- seq(from = 40, to = 100, length.out = 500)
yy.tn <- sapply(xx, function(x) CTMI(x, nlm.tn$estimate))
lines(xx, yy.tn)

```



Code for figure 6.11 page 120.

```
x.PCC7941 <- c(20, 25, 27.5, 32.5, 35)
y.PCC7941 <- c(0.58, 0.67, 1.05, 1.16, 1.01)
x.CYA140 <- c(20, 25, 27.5, 30, 32.5, 35)
y.CYA140 <- c(0.26, 0.77, 0.82, 0.94, 0.93, 0.70)
plot(x.PCC7941, y.PCC7941, las = 1,
     xlab = "Temperature [°C]", ylab = "Specific growth rate [1/day]",
     main = "Growth rate as a function of temperature", xlim = c(0, 50),
     ylim = c(0, 1.2))
points(x.CYA140, y.CYA140, pch = 19, col = "red")
nlm.PCC7941 <- nlm.CTMI(c(10, 30, 40, 1.2), cbind(x.PCC7941, y.PCC7941))
nlm.CYA140 <- nlm.CTMI.auto(cbind(x.CYA140, y.CYA140))
xx <- seq(from = 0, to = 50, length.out = 500)
yy.PCC7941 <- sapply(xx, function(x) CTMI(x, nlm.PCC7941$estimate))
yy.CYA140 <- sapply(xx, function(x) CTMI(x, nlm.CYA140$estimate))
lines(xx, yy.PCC7941)
lines(xx, yy.CYA140, col = "red")
legend("topleft", inset = 0.02, pch = c(1, 19), col = c("black", "red"),
      legend = c("PCC7941", "CYA140"))
```



Code for figure 6.15 page 140.

```
pK <- c(7.555, 3.872, 4.4412, 5.637, 9.052, 11.84, 10.85)
names(pK) <- s2c("CDEHKRY")
computePositiveCharge <- function(aa, pH, pK) {
  10^(-pH)/(10^(-pK[aa]) + 10^(-pH))
}
xx <- seq(0, 14, length.out = 255)
yy <- sapply(xx, computePositiveCharge, aa = "R", pK = pK)
plot(xx, yy, type = "l", las = 1, ylab = "Positive charge", xlab = "pH",
     main = "Positively charged aminoacids")
yy <- sapply(xx, computePositiveCharge, aa = "K", pK = pK)
lines(xx, yy, lty = 2, col = "darkblue")
yy <- sapply(xx, computePositiveCharge, aa = "H", pK = pK)
lines(xx, yy, lty = 3, col = "red")
legend("bottomleft", inset = 0.01, legend = c("Arg (R)", "Lys (K)", "His (H)"),
      lty = 1:3, col = c("black", "darkblue", "red"))
```



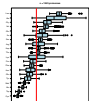
Code for figure 6.16 page 141.

```
computeNegativeCharge <- function(aa, pH, pK) {
  10^(-pK[aa])/(10^(-pK[aa]) + 10^(-pH))
}
xx <- seq(0, 14, length.out = 255)
yy <- sapply(xx, computeNegativeCharge, aa = "D", pK = pK)
plot(xx, yy, type = "l", las = 1, ylab = "Negative charge (absolute value)", xlab = "pH",
     main = "Negatively charged aminoacids")
yy <- sapply(xx, computeNegativeCharge, aa = "E", pK = pK)
lines(xx, yy, lty = 2, col = "darkblue")
yy <- sapply(xx, computeNegativeCharge, aa = "C", pK = pK)
lines(xx, yy, lty = 3, col = "red")
yy <- sapply(xx, computeNegativeCharge, aa = "Y", pK = pK)
lines(xx, yy, lty = 4, col = "purple")
legend("topleft", inset = 0.01, legend = c("Asp (D)", "Glu (E)", "Cys (C)", "Tyr (Y)"),
      lty = 1:4, col = c("black", "darkblue", "red", "purple"))
```



Code for figure 6.17 page 142.

```
computeCharge <- function(pH, prot, pK){
  faa <- table(factor(s2c(prot), levels = s2c("RKHDECY")))
  pos <- faa["R"]*computePositiveCharge("R", pH, pK) +
    faa["K"]*computePositiveCharge("K", pH, pK) +
    faa["H"]*computePositiveCharge("H", pH, pK)
  neg <- faa["D"]*computeNegativeCharge("D", pH, pK) +
    faa["E"]*computeNegativeCharge("E", pH, pK) +
    faa["C"]*computeNegativeCharge("C", pH, pK) +
    faa["Y"]*computeNegativeCharge("Y", pH, pK)
  return(pos - neg)
}
xx <- seq(0, 14, length.out = 255)
yy <- sapply(xx, computeCharge, prot = "RKHDECY", pK = pK)
plot(xx, yy, type = "l", las = 1, ylab = "Charge", xlab = "pH",
     main = "RKHDECY")
abline(h = 0, lty = 2)
pI <- uniroot(computeCharge, c(0,14), prot = "RKHDECY", pK = pK)$root
abline(v = pI, lty = 2, col = "red")
```



Code for figure 6.14 page 139.

```
iAla <- which(colnames(tdd) == "Ala")
tddaa <- tdd[, iAla:(iAla + 20)]
tddaa <- tddaa[, ~which(colnames(tddaa) == "Stp")]
boxplot(tddaa[, , order(colMeans(tddaa))], horizontal = TRUE, las = 1,
       xlab = "Aminoacid frequency [X]", col = "lightblue",
       main = paste("n =", nrow(tddaa), "proteomes"))
abline(v = 5, lty = 2, col = "red")
```



Code for figure 6.12 page 136.

```
TILS2002 <- read.table("local/TILS2002.csv", sep = "\t", header = TRUE)
x <- TILS2002$GC
y <- 100*TILS2002$GCIGR
main <- paste("Intergenic GC content\nn =", nrow(TILS2002), "chromosomes")
plot(x, y, xlab = "Genomic GC content [M]", las = 1,
      ylab = "Intergenic GC content [K]", main = main)
abline(c(0, 1), lty = 2)
lm1 <- lm(y~x)
abline(lm1)
legend("topleft", inset = 0.01, legend = c("y = x", "linear fit"),
      lty = 2:1)
mtext(paste("Slope =", signif(lm1$coef[2], 4)), adj = 0)
mtext(paste("Intercept =", signif(lm1$coef[1], 4)), adj = 1)
r2 <- signif(cor(x, y)^2, 3)
mtext(bquote(r^2 == .(r2)), adj = 0.5)
iout <- 47
points(x[iout], y[iout], pch = 19)
```



Code for figure 6.18 page 144.

```
main <- paste("Isoelectric point distribution\nn =", nrow(tdd), "proteomes")
plot(density(tdd$pI, adjust = 0.5), main = main,
      xlab = "pI", las = 1)
col <- rep("black", nrow(tdd))
col[!is.na(tdd$class) & tdd$class == 183963] <- "red"
rug(tdd$pI)
rug(tdd$pI[col=="red"], col = "red")
legend("topright", inset = 0.02, lwd = 1, legend = c("Halobacteria", "Other"),
      col = c("red", "black"))
```

8.2 Session information

```
sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Sierra 10.12.6
Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib

locale:
[1] fr_FR.UTF-8/fr_FR.UTF-8/fr_FR.UTF-8/C/fr_FR.UTF-8/fr_FR.UTF-8

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

other attached packages:
[1] xtable_1.8-3 MASS_7.3-50 seqinr_3.4-5 ade4_1.7-13

loaded via a namespace (and not attached):
[1] compiler_3.5.1 tools_3.5.1
```


The file `backmatter` is empty.

Bibliography

- [1] S. Aksoy. *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., taxa consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies. *International Journal of Systematic Bacteriology*, 45:848–851, 1995.
- [2] A.C. Astwood and A.C. Wais. Psychrotrophic bacteria isolated from a constantly warm tropical environment. *Curr. Microbiol.*, 36:148–151, 1998.
- [3] H. Atomi, T. Fukui, T. Kanai, M. Morikawa, and T. Imanaka. Description of *Thermococcus kodakaraensis* sp. nov., a well studied hyperthermophilic archaeon previously reported as *Pyrococcus* sp. KOD1. *Archaea*, 1:263–267, 2004.
- [4] H.L. Ayala-del Río, P.S. Chain, J.J. Grzymiski, M.A. Ponder, N. Ivanova, P.W. Bergholtz, G. Di Bartolo, L. Hauser, M. Land, C. Bakermans, D. Rodrigues, J. Klappenbach, D. Zarka, F. Larimer, P. Richardson, A. Murray, M. Thomashow, and J.M. Tiedje. The genome sequence of *Psychrobacter arcticus* 273-4, a psychroactive Siberian permafrost bacterium reveals mechanisms for adaptation to low temperature growth. *Appl. Environ. Microbiol.*, 76:2304–2312, 2010.
- [5] S.S. Bae, Y.J. Kim, S.H. Yang, J.K. Lim, J.H. Jeon, H.S. Lee, S.G. Kang, S.-J. Kim, and J.-H. Lee. *Thermococcus onnurineus* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent area at the PAC-MANUS field. *Journal of Microbiology and Biotechnology*, 16:1826–1831, 2006.
- [6] C. Bakermans, L. Ayala-del Río, Ponder M.A., T. Vishnivetskaya, D. Gilichinsky, M.F. Thomashow, and J.M. Tiedje. *Psychrobacter cryohalolentis* sp. nov. and *Psychrobacter arcticus* sp. nov., isolated from siberian permafrost. *International Journal of Systematic and Evolutionary Microbiology*, 56:1285–1291, 2006.
- [7] M. Banerjee, R.C. Everroad, and R.W. Castenholz. Studies on *Aphanotheca halophytica* FRÉMY from a solar pond: comparison of two isolates on the basis of cell polymorphism and growth response to salinity, temperature and light conditions. *Bot. Mar.*, 28:389–398, 1985.
- [8] M. Banerjee, R.C. Everroad, and R.W. Castenholz. An unusual cyanobacterium from saline thermal waters with relatives from unexpected habitats. *Extremophiles*, 13:707–716, 2009.

- [9] M.A. Barber. The rate of multiplication of *Bacillus coli* at different temperatures. *Journal of Infectious diseases*, 5:379–400, 1908.
- [10] S. Belkin, C.O. Wirsen, and H.W. Jannasch. A new sulfur-reducing, extremely thermophilic eubacterium from a submarine thermal vent. *Applied and Environmental Microbiology*, 51:1180–1185, 1986.
- [11] A.N. Belozersky and A.S. Spirin. A correlation between the compositions of deoxyribonucleic and ribonucleic acids. *Nature*, 182:111–112, 1958.
- [12] O. Bernard and B. Rémond. Validation of a simple model accounting for light and temperature effect on microalgal growth. *Bioresource Technology*, 123:520–527, 2012.
- [13] L.R. Beuchat. Environmental factors affecting survival and growth of *Vibrio parahaemolyticus*. a review. *J. Milk Food Technol.*, 38:476–480, 1975.
- [14] B. Bjellqvist, G.J. Hughes, C. Pasquali, N. Paquet, F. Ravier, J.-C. Sanchez, S. Frutiger, and D.F. Hochstrasser. The focusing positions of polypeptides in immobilized ph gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14:1023–1031, 1993.
- [15] G.F. Blanchard, J.M. Guarini, P. Richard, P. Gros, and F. Mornet. Quantifying the short-term temperature effect on light-saturated photosynthesis of intertidal microphytobenthos. *Marine Ecology Progress Series*, 134:309–313, 1996.
- [16] D.J. Brenner, F.J. Fanning, G.R. and Skerman, and S. Falkow. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *Journal of Bacteriology*, 109:953–965, 1972.
- [17] T.D. Brock and H. Freeze. *Thermus aquaticus* gen. n. and sp. n., a non-sporulating extreme thermophile. *Journal of Bacteriology*, 98:289–297, 1969.
- [18] S.D. Brown, M.B. Begemann, M.R. Mormile, J.D. Wall, C.S. Han, L.A. Goodwin, S. Pitluck, M.L. Land, L.J. Hauser, and D.A. Elias. Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium *Halanaerobium hydrogeniformans*. *Journal of Bacteriology*, 193:3682–3683, 2011.
- [19] R.E. Buchanan. Life phases in a bacterial culture. *Journal of Infectious Diseases*, 23:109–125, 1918.
- [20] D.G. Burns, P.H. Janssen, T. Itoh, M. Kamekura, A. Echigo, and M.L. Dyll-Smith. *Halonotius pteroides* gen. nov., sp. nov., an extremely halophilic archaeon recovered from a saltern crystallizer. *International Journal of Systematic and Evolutionary Microbiology*, 60:1196–1199, 2010.
- [21] D.G. Burns, P.H. Janssen, T. Itoh, M. Kamekura, Z. Li, F. Jensen, G. Rodríguez-Valera, H. Bolhuis, and M.L. Dyll-Smith. *Haloquadratum walsbyi* gen. nov., sp. nov., the square haloarchaeon of Walsby, isolated from saltern crystallizers in Australia and Spain. *International Journal of Systematic and Evolutionary Microbiology*, 57:387–392, 2007.

- [22] J.-L. Cayol, B. Ollivier, B.K.C. Patel, G. Prensier, J. Guezennec, and J.-L. Garcia. Isolation and characterization of *Halothemotrix orenii* gen. nov., sp. nov., a halophilic, thermophilic, fermentative, strictly anaerobic bacterium. *International Journal of Systematic Bacteriology*, 44:534–540, 1994.
- [23] D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In H.E. Roman U. Bastolla, M. Porto and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, USA, 2007. ISBN 978-3-540-35305-8.
- [24] R. Corkrey, T.A. McMeekin, J.P. Bowman, D.A. Ratkowsky, J. Olley, and T. Ross. Protein thermodynamics can be predicted directly from biological growth rates. *PloS one*, 9:e96, 2014.
- [25] J.Z. Dalggaard and A. Garrett. *The biochemistry of Archaea (Archaeobacteria)*, chapter Archaeal hyperthermophile genes. Elsevier Science, Amsterdam, 1993.
- [26] R. de Levie. The Henderson-Hasselbalch equation: Its history and limitations. *Journal of Chemical Education*, 80:146–146, 2003.
- [27] M. De Rosa, A. Gambacorta, and J.D. Bu’lock. Exteremly thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius*. *Journal of General Microbiology*, 86:156–164, 1975.
- [28] A.C.R. Dean and P.L. Rogers. The cell size and macromolecular composition of *Aerobacter aerogenes* in various sytems of continuous culture. *Biochimica Biophysica Acta*, 148:267–279, 1967.
- [29] K.A. Dill, K. Ghosh, and J.D. Schmit. Physical limits of cells and proteomes. *Proceedings of the national academy of sciences of the United States of America*, 108:17876–17882, 2011.
- [30] I.N. Dominova, I.V. Kublanov, O.A. Podosokorskaya, K.S. Derbikova, M.V. Patruchev, and S.V. Toshchakov. Complete genomic sequence of “*Thermofilum adornatus*” strain 1910b^T, a hyperthermophilic anaerobic organotrophic crenarchaeon. *Genome Announc.*, 5:e00726–13, 2013.
- [31] M.K.M. Engqvist. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC Microbiology*, 18:177, 2018.
- [32] G. Erauso, A.-L. Reysenbach, A. Godfroy, J.-R. Meunier, B. Crump, F. Partensky, J.A. Baross, V. Marteinsson, G. Barbier, N.R. Pace, and D. Prieur. *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *Archives of Microbiology*, 160:338–349, 1993.

- [33] C.R. Everroad, H. Otaki, K. Matsuura, and S. Haruta. Diversification of bacterial community composition along a temperature gradient at a thermal spring. *Microbes Environ.*, 27:374–381, 2012.
- [34] J.G. Fox, F.E. Dewhirst, J.G. Tully, B.J. Paster, L. Yan, N.S. Taylor, M.J. Collins, P.L. Gorelick, and J.M. Ward. *Helicobacter hepaticus* sp.nov., a microaerophilic bacterium isolated from livers and intestinal mucosal scrapings from mice. *Journal of Clinical Microbiology*, 32:1238–1245, 1994.
- [35] C.M. Fraser, S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum, E. Sodergren, J.M. Hardham, M.P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J.K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M.D. Cotton, C. Fujii, S. Garland, B. Hatch, K. Horst, K. Roberts, M. Sandusky, J. Weidman, H.O. Smith, and J.C. Venter. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, 281:375–388, 1998.
- [36] V.A. Gaisin, A.M. Kalashnikov, D.S. Grouzdev, M.V. Sukhacheva, B.B. Kuznetsov, and V.M. Gorlenko. *Chloroflexus islandicus* sp. nov., a thermophilic filamentous anoxygenic phototrophic bacterium from a geyser. *International Journal of Systematic and Evolutionary Microbiology*, 67:1381–1386, 2017.
- [37] M.Y. Galperin, V. Brover, I. Tolstoy, and N. Yutin. Phylogenomic analysis of the family *Peptostreptococcaceae* (*Clostridium* cluster XI) and proposal for reclassification of *Clostridium litorale* (FENDRICH *et al.* 1991) and *Eubacterium acidaminophilum* (ZINDEL *et al.* 1989) as *Peptoclostridium litorale* gen. nov. comb. nov. and *Peptoclostridium acidaminophilum* comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, 66:5506–5513, 2016.
- [38] N. Galtier and J.R. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, 44:632–635, 1997.
- [39] A. Gorlas, O. Croce, J. Oberto, E. Gaudiard, P. Forterre, and E. Marguet. *Thermococcus nautili* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep-sea vent. *International Journal of Systematic and Evolutionary Microbiology*, 64:1802–1810, 2014.
- [40] Student [Gosset, W.S.]. The probable error of a mean. *Biometrika*, 6:1–25, 1908.
- [41] R. Grantham, C. Gautier, M. Gouy, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research*, 9:r43–r74, 1981.
- [42] C.J. Griffith and T.H. Melville. Growth of oral streptococci in a chemostat. *Archives of Oral Biology*, 19:87–90, 1974.
- [43] G.M. Grimaud. *Modelling the temperature effect on phytoplankton: from acclimation to adaptation*. PhD thesis, Université Nice Sophia-Antipolis, France, 2016.

- [44] G.M. Grimaud, F. Mairet, A. Sciandra, and O. Bernard. Modeling the temperature effect on the specific growth rate of phytoplankton: a review. *Reviews in Environmental Science and Bio/Technology*, 16:625–645, 2017.
- [45] D.W. Grogan. Phenotypic characterization of the archaeobacterial genus *Sulfolobus*: Comparison of five wild-type strains. *Journal of bacteriology*, 171:6710–6719, 1989.
- [46] M.-H. Guinebrière, F.L. Thompson, A. Sorokin, P. Normand, P. Dawyndt, M. Ehling-Schulz, B. Svensson, V. Sanchis, C. Nguyen-The, M. Heyndrickx, and P. De Vos. Ecological diversification in the *Bacillus cereus* group. *Environmental Microbiology*, 10:851–865, 2008.
- [47] K.A. Hasselbalch. Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebundenen Kohlensäure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl. *Biochemische Zeitschrift*, 78:112–144, 1917.
- [48] L.J. Henderson. Concerning the relationship between the strength of acids and their capacity to preserve neutrality. *Am. J. Physiol.*, 21:173–179, 1908.
- [49] C.N. Hinshelwood. *The chemical kinetics of the bacterial cell*, chapter Influence of temperature on the growth of bacteria, pages 254–257. Clarendon Press, Oxford, UK, 1946.
- [50] S.-L. Huang, L.-C. Wu, H.-K. Liang, K.-T. Pan, J.-T. Horng, and M.-T. Ko. PGTdb: a database providing growth temperatures of prokaryotes. *Bioinformatics*, 20:276–278, 2004.
- [51] G. Jiménez, M. Urdiain, A. Cifuentes, A. López-López, A.R. Blanch, J. Tamames, P. Kämpfer, A.-B. Kolstø, D. Ramón, J.F. Martínez, F.M. Codoñer, and R. Rosselló-Móra. Description of *Bacillus toyonensis* sp. nov., a novel species of the *Bacillus cereus* group, and pairwise genome comparisons of the species of the group by means of ANI calculations. *Systematic and Applied Microbiology*, 36:383–391, 2013.
- [52] F.H. Johnson and I. Lewin. The growth rate of *E. coli* in relation to temperature, quinine and coenzyme. *Journal of Cellular and Comparative Physiology*, 28:47–75, 1946.
- [53] M. Kahn. An exhalent problem for teaching statistics. *Journal of Statistical Education*, 13(2), 2005.
- [54] C. Kato and Y. Nogi. Correlation between phylogenetic structure and function: examples from deep-sea *Shewanella*. *FEMS Microbiology Ecology*, 35:223–230, 2001.
- [55] C. Kato, T. Sato, and K. Horikoshi. Isolation and properties of barophilic and barotolerant bacteria from deep-sea mud samples. *Biodiversity and Conservation*, 4:1–9, 1995.

- [56] S.P. Kennedy, W.V. Ng, S.T. Salzberg, L. Hood, and S. DasSarma. Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research*, 11:1641–1650, 2001.
- [57] D. Kim, J. Chun, Y.C. Sahin, N. Hah, and M. Goodfellow. Analysis of thermophilic clades within the genus *Streptomyces* by 16S ribosomal DNA sequence comparisons. *International Journal of Systematic Bacteriology*, 46:581–587, 1996.
- [58] A.L. Koch. Turbidity measurement of bacterial cultures in some available commercial instruments. *Analytical Biochemistry*, 38:252–259, 1970.
- [59] T.V. Kochetkova, I.I. Rusanov, N.V. Pimenov, T.V. Kolganova, A.V. Lebedinsky, E.A. Bonch-Osmolovskaya, and T.G. Sokolova. Anaerobic transformation of carbon monoxide by microbial communities of Kamchatka hot springs. *Extremophiles*, 15:319–325, 2011.
- [60] S.A.L.M. Kooijman. *Dynamic energy budget theory for metabolic organisation*. Cambridge university press, UK, 2010.
- [61] G. Korkmaz, M. Holm, T. Wiens, and S. Sanyal. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *The Journal of Biological Chemistry*, 249:3034–3042, 2014.
- [62] L.P. Kozłowski. IPC - isoelectric point calculator. *Biology Direct*, 11:55, 2016.
- [63] J.K. Kristjánsson, S. Hjörleifsdóttir, V.T. Marteinsson, and G.A. Alfredsson. *Thermus scotoductus*, sp. nov., a pigment-producing thermophilic bacterium from hot tap water in iceland and including *Thermus* sp. X-1. *Systematic and Applied Microbiology*, 17:44–50, 1994.
- [64] Lee K.Y., R. Wahl, and E. Barbu. Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries. *Annales de l'Institut Pasteur*, 91:212–224, 1956.
- [65] J. Lalucat, A. Bennasar, R. Bosch, E. García-Valdés, and N.J. Palleroni. Biology of *Pseudomonas stutzeri*. *Microbiology and Molecular Biology Reviews*, 70:510–547, 2006.
- [66] J.K. Lanyi. Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriological Reviews*, 38:272–290, 1974.
- [67] M. le Roes and P.R. Meyers. *Streptomyces pharetrae* sp. nov., isolated from soil from the semi-arid Karoo region. *Systematic and Applied Microbiology*, 28:488–493, 2005.
- [68] M. Lee and A.C. Chandler. A study of the nature, growth and control of bacteria in cutting compounds. *Journal of Bacteriology*, 41:373–386, 1941.
- [69] J.R. Lobry. *Ré-évaluation de modèle de croissance de Monod. Effet des antibiotiques sur l'énergie de maintenance*. PhD thesis, University Claude Bernard, Lyon I, 1991.

- [70] J.R. Lobry. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, 205:309–316, 1997.
- [71] J.R. Lobry. *Multivariate Analyses of Codon Usage*. ISTE Editions, 27-37 St George’s Road, Londres SW19 4EU, United Kingdom, 2018.
- [72] J.R. Lobry and C. Gautier. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, 22:3174–3180, 1994.
- [73] J.R. Lobry and A. Necşulea. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, 385:128–136, 2006.
- [74] J.R. Lobry, L. Rosso, and J.-P. Flandrois. A FORTRAN subroutine for the determination of parameter confidence limits in non-linear models. *Binary*, 3:86–93, 1991.
- [75] J.R. Lobry and N. Sueoka. Asymmetric directional mutation pressures in bacteria. *Genome Biology*, 3(10):research0058.1–research0058.14, 2002.
- [76] R. Luedeking and E.L. Piret. A kinetic study of the lactic acid fermentation. *Journal of Biochemical and Microbiological Technology and Engineering*, 1:393–412, 1959.
- [77] M. Lüring, F. Eshetu, E.J. Faassen, S. Kosten, and V.L.M. Huszar. Comparison of cyanobacterial and green algal growth rates at different temperatures. *Freshwater Biology*, 58:1–8, 2012.
- [78] M. Mahmoudi, A.A. Arab, J. Zahiri, and Y. Parandian. An overview of the protein thermostability prediction: Databases and tools. *J. Nanomed. Res.*, 3:0007, 2016.
- [79] M. Mandel and J. Marmur. [109] use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. *Methods in Enzymology*, 12:195–206, 1968.
- [80] A.V. Mardanov, T.V. Kochetkova, A.V. Beletsky, E.A. Bonch-Osmolovskaya, N.V. Ravin, and K.G. Skryabin. Complete genome sequence of the hyperthermophilic cellulolytic crenarchaeon “*Thermogladius cellulolyticus*” 1633. *Journal of Bacteriology*, 194:4446–4447, 2012.
- [81] D. Maurin and D. Raoult. Current knowledge of *Bartonella* species. *Eur. J. Clin. Microbiol. Infect. Dis.*, 16:487–506, 1997.
- [82] S.R. Miller and R.W. Castenholz. Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechococcus*. *Applied and Environmental Microbiology*, 66:4222–4229, 2000.
- [83] A. Mira, H. Ochman, and N.A. Moran. Deletional bias and the evolution of bacterial genomes. *TRENDS in Genetics*, 17:589–596, 2001.
- [84] J. Monod. *Recherches sur la croissance des cultures bactériennes*. PhD thesis, Paris, 1941.

- [85] R.L. Moore, R.M. Weiner, and R. Gebers. Genus *Hyphomonas* Pongratz 1957 norn. rev. emend. *Hyphomonas polymorpha* Pongratz 1957 norn. rev. emend. and *Hyphomonas neptunium* (Leifson 1964) comb. nov. emend. (*Hyphomicrobium neptunium*. *International Journal of Systematic Bacteriology*, 34:71–73, 1984.
- [86] M. Morikawa, Y. Izawa, N. Rashid, T. Hoaki, and T. Imanaka. Purification and characterization of a thermostable thiol protease from a newly isolated hyperthermophilic *Pyrococcus* sp. *Applied and Environmental Microbiology*, 60:4559–4566, 1994.
- [87] M.A. Munson, P. Baumann, and M.G. Kinsey. *Buchnera* gen. nov. and *Buchnera aphidicola* sp. nov., a taxon consisting of the mycetocyte-associated, primary endosymbionts of aphids. *International Journal of Systematic Bacteriology*, 41:566–568, 1991.
- [88] N.C.S. Mykytczuk, S.J. Foote, G. Southam, C.W. Greer, and L.G. Whyte. Bacterial growth at -15°C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1. *The ISME Journal*, 2013:1–16, 2013.
- [89] L.K. Nakamura, I. Blumenstock, and D. Claus. Taxonomic study of *Bacillus coagulans* hammer 1915 with a proposal for *Bacillus smithii* sp. nov. *International Journal of Systematic Bacteriology*, 38:63–73, 1988.
- [90] H. Naya, H. Romero, A. Zavala, B. Alvarez, and H. Musto. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *Journal of Molecular Evolution*, 55:260–264, 2002.
- [91] T.D. Niederberger, D.K. Götz, I.R. McDonald, R.S. Ronimus, and H.W. Morgan. *Ignisphaera aggregans* gen. nov., sp. nov., a novel hyperthermophilic crenarchaeote isolated from hot springs in Rotorua and Tokaanu, New Zealand. *International Journal of Systematic and Evolutionary Microbiology*, 56:965–971, 2006.
- [92] Y. Nogi, C. Kato, and K. Horikoshi. Taxonomic studies of deep-sea barophilic *Shewanella* strains and description of *Shewanella violacea* sp. nov. *Arch. Microbiol.*, 170:331–338, 1998.
- [93] J. Norberg. Biodiversity and ecosystem functioning: A complex adaptive systems approach. *Limnol. Oceanogr.*, 49:1269–1277, 2004.
- [94] J. Oezle and R.C. Fuller. Temperature dependence of growth and membrane-bound activities of *Chloroflexus aurantiacus* energy metabolism. *Journal of Bacteriology*, 155:90–96, 1983.
- [95] A. Oren. *The prokaryotes. A handbook on the biology of bacteria: ecophysiology, isolation, identification, applications, 3rd ed.*, chapter The order Halobacteriales. Springer-Verlag, New York, N.Y., 2001.
- [96] A. Oren. Naming cyanophyta/cyanobacteria - a bacteriologist's view. *Fottea*, 11:9–16, 2011.

- [97] T. Oshima and K. Imahori. Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. *International Journal of Systematic Bacteriology*, 24:102–112, 1974.
- [98] E.L. Overholser and R.H. Taylor. The vegetation of the bay of Fundy salt and diked marshes: an ecological study. *The botanical gazette*, XXXVI:429–455, 1903.
- [99] E.L. Overholser and R.H. Taylor. Ripening of pears and apples as modified by extreme temperatures. *The botanical gazette*, LXIX:273–296, 1920.
- [100] B. Pace and L.L. Campbell. Correlation of maximal growth temperature and ribosome heat stability. *Proceedings of the national academy of sciences of the United States of America*, 57:1110–1116, 1967.
- [101] V. Palanichamy, A. Hundet, B. Mitra, and N. Reddy. Optimization of cultivation parameters for growth and pigment production by *Streptomyces* spp. isolated from marine sediment and rhizosphere soil. *International Journal of Plant, Animal and Environmental Sciences*, 1:158–170, 2011.
- [102] S. Penel, A.M. Arigon, J.F. Dufayard, A.S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10:S3, 2009.
- [103] B.K. Pierson and R.W. Castenholz. A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen. and sp. nov. *Arch. Microbiol.*, 100:5–24, 1974.
- [104] J. Pittera, F. Humily, M. Thorel, D. Grulois, L. Garczarek, and C. Six. Connecting thermal physiology and latitudinal niche partitioning in marine *Synechococcus*. *The ISME Journal*, 8:1221–1236, 2014.
- [105] H.N. Po and N.M. Senozan. The Henderson-Hasselbalch equation: Its history and limitations. *Journal of Chemical Education*, 78:1499–1503, 2001.
- [106] S. Podell, J.A. Ugalde, P. Narasingarao, J.F. Banfield, K.B. Heidelberg, and E.E. Allen. Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLoS ONE*, 8:e61692, 2013.
- [107] J.S. Poindexter. Biological properties and classification of the *Caulobacter* group. *Bacteriological Reviews*, 28:231–295, 1964.
- [108] I.S. Povolotskaya, F.A. Kondrashov, A. Ledda, and P.K. Vlasov. Stop codons in bacteria are not selectively equivalent. *Biology Direct*, 7:30, 2012.
- [109] N. Rashid, M. Morikawa, K. Nagahisa, S. Kanaya, and T. Imanaka. Characterization of a RecA/RAD51 homologue from the hyperthermophilic archaeon *Pyrococcus* sp. KOD1. *Nucleic Acids Research*, 25:719–726, 1997.
- [110] D.A. Ratkowsky, R.K. Lowry, T.A. McMeekin, A.N. Stokes, and R.E. Chandler. Model for bacterial culture growth rate throughout the entire biokinetic range. *Journal of Bacteriology*, 154:1222–1226, 1983.

- [111] D.A. Ratkowsky and G.V.P. Reddy. Empirical model with excellent statistical properties for describing temperature-dependent developmental rates of insects and mites. *Annals of the Entomological Society of America*, 110:302–309, 2017.
- [112] Margheri, M.C. and Ventura, S. and Kaštovský, J. and Komárek, J. The taxonomic validation of the cyanobacterial genus *Halotheca*. *Phycologia*, 47:477–486, 2008.
- [113] Rainey, F.A. and Hollen, B.J. and Small, A. *BERGEY'S Manual of Systematic Bacteriology, 2nd edition, Volume 3: The Firmicutes*, chapter Genus I. Clostridium, pages 738–828. Springer, New York, USA, 2009.
- [114] E.R. Reichenberger, G. Rosen, U. Hershberg, and R. Hershberg. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biology and Evolution*, 7:1380–1389, 2015.
- [115] J.L. Robinson, B. Pyzyna, R.G. Atrasz, C.A. Henderson, K.L. Morrill, A.M. Burd, E. DeSoucy, R.E. Fogleman III, J.B. Naylor, S.M. Steele, D.R. Elliot, K.J. Leyva, and R.F. Shand. Growth kinetics of extremely halophilic *Archaea* (family *Halobacteriaceae*) as revealed by Arrhenius plots. *Journal of Bacteriology*, 187:923–929, 2005.
- [116] R. Rolfe and M. Meselson. The relative homogeneity of microbial DNA. *Proceedings of the national academy of sciences of the United States of America*, 45:1039–1043, 1959.
- [117] B. Rosner. *Fundamentals of Biostatistics*. PWS-Kent, Boston, Massachusetts, 1990.
- [118] L. Rosso, J.R. Lobry, and J.-P. Flandrois. An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model. *Journal of Theoretical Biology*, 162(4):447–463, 1993.
- [119] D.W. Roush. Production of 1,3-propanediol from glycerol under haloalkaline conditions by *Halanaerobium hydrogeniformans*. Master's thesis, Missouri University of Science and Technology, 2013.
- [120] C.E. Safford, J.M. Sherman, and H.M. Hodge. *Streptococcus salivarius*. *Journal of Bacteriology*, 33:263–274, 1937.
- [121] R. Saha, C. Spröer, B. Beck, and S. Bagley. *Pseudomonas oleovorans* subsp. *lubricantis* subsp. nov., and reclassification of *Pseudomonas pseudoalcaligenes* ATCC 17440^T as later synonym of *Pseudomonas oleovorans* ATCC 8062^T. *Curr. Microbiol.*, 60:294–300, 2010.
- [122] M Schaechter, O. Maaløe, and N.O. Kleldgaard. Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonella typhimurium*. *Journal of General Microbiology*, 19:592–606, 1958.
- [123] L.A. Schipper, J.K. Hobbs, S. Ruledge, and V.L. Arcus. Thermodynamic theory explains the temperature optima of soil microbial processes and high q_{10} values at low temperatures. *Global Change Biology*, 20:3578–3586, 2014.

- [124] C. Schleper, G. Pühler, H.-P. Klenk, and W. Zillig. *Picrophilus oshimae* and *Picrophilus torridus* fam. nov., gen. nov., sp. nov., two species of hyperacidophilic, thermophilic, heterotrophic, aerobic archaea. *International Journal of Systematic Bacteriology*, 46:814–816, 1996.
- [125] E. Scolnick, R. Tompkins, T. Caskey, and M. Nirenberg. Release factors differing in specificity for terminator codons. *Proceedings of the national academy of sciences of the United States of America*, 61:768–774, 1968.
- [126] B. Shi and X. Xia. Changes in growth parameters of *Pseudomonas pseudoalcaligenes* after ten months culturing at increasing temperature. *FEMS Microbiology Ecology*, 45:127–134, 2003.
- [127] E.B. Shirling and D. Gottlieb. Cooperative description of the type cultures of *Streptomyces*. II. species descriptions from first study. *International Journal of Systematic Bacteriology*, 18:69–189, 1968.
- [128] D.Y. Sorokin, I.V. Kublanov, S.N. Gavrillov, D. Rojo, P. Roman, P.N. Golyshin, V.Z. Slepak, F. Smedile, M. Ferrer, E. Messina, V. La Cono, and M.M. Yakimov. Elemental sulfur and acetate can support life of a novel strictly anaerobic haloarchaeon. *The ISME Journal*, 10:240–252, 2016.
- [129] D.Y. Sorokin, I.V. Kublanov, M.M. Yakimov, W.I.C. Rijpstra, and J.S.S. Damsté. *Halanaeroarchaeum sulfurireducens* gen. nov., sp. nov., the first obligately anaerobic sulfur-respiring haloarchaeon, isolated from a hypersaline lake. *International Journal of Systematic and Evolutionary Microbiology*, 66:2377–2381, 2016.
- [130] J. Spaun. Problems in standardization of turbidity determinations of bacterial suspensions. *Bulletin of the World Health Organisation*, 26:219–255, 1962.
- [131] T. Stadtman and H.A. Barker. Studies on the methane fermentation. X. a new formate-decomposing bacterium, *Methanococcus vannielii*. *Journal of Bacteriology*, 62:269–280, 1951.
- [132] T. Stadtman and L.S. McClung. *Clostridium sticklandii* nov. spec. *Journal of Bacteriology*, 73:218–219, 1957.
- [133] J.T. Staley, M.P. Bryant, N. Pfennig, and J.G. Holt. *Bergey's manual of systematic bacteriology*. Williams and Wilkins, Baltimore, 1984.
- [134] S. Stolyar, Z. Liu, V. Thiel, L.P. Tomsho, N. Pinel, W.C. Nelson, S.R. Lindemann, M.F. Romine, S. Haruta, S.C. Schuster, D.A. Bryant, and J.K. Fredrickson. Genome sequence of the thermophilic cyanobacterium *Thermosynechococcus* sp. strain NK55a. *Genome Announc.*, 2:e01060–13, 2014.
- [135] N. Sueoka. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proceedings of the national academy of sciences of the United States of America*, 45:1480–1490, 1959.

- [136] N. Sueoka. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1141–1149, 1961.
- [137] N. Sueoka. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, 48:582–592, 1962.
- [138] N. Sueoka, J. Marmur, and P. Doty. Heterogeneity in deoxyribonucleic acids. II. dependence of the density of deoxyribonucleic acids on guanine-cytosine. *Nature*, 183:1427–1431, 1959.
- [139] Sun Microsystems. XDR: external data representation standard. RFC 1014. Technical report, Network Working Group, 1987.
- [140] C. Söhnngen, A. Podstawka, B. Bunk, D. Gleim, A. Vetcinina, L.C. Reimer, C. Ebeling, C. Pendarovski, and J. Overmann. BacDive - the bacterial diversity metadatabase in 2016. *Nucleic Acids Research*, 44:D581–D585, 2016.
- [141] I.B. Tager, B. Weiss, S.T. and Rosner, and F. E. Speizer. Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology*, 110:15–26, 1979.
- [142] Y. Takahata, M. Nishijima, T. Hoaka, and T. Maruyama. *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan. *International Journal of Systematic and Evolutionary Microbiology*, 51:1901–1909, 2001.
- [143] M.L. Tamplin, R. Phillips, T.A. Stewart, J.B. Luchansky, and L.C. Kelley. Behavior of *Bacillus anthracis* strains Sterne and Ames K0610 in sterile raw ground beef. *Applied and Environmental Microbiology*, 74:1111–1116, 2008.
- [144] S.M. Tiquia-Arashiro. *Thermophilic Carboxydrotrophs and their Applications in Biotechnology*. Springer, New York, U.S.A., 2014.
- [145] G. Toennies and D.L. Gallant. The relation between photometric turbidity and bacterial concentration. *Growth*, 13:7–20, 1949.
- [146] S.V. Toshchakov, A.A. Korzhenkov, N.I. Samarov, I.O. Mazunin, O.I. Mozhey, I.S. Shmyr, K.S. Derbikova, E.A. Taranov, I.N. Dominova, E.A. Bonch-Osmolovskaya, M.V. Patrushev, O.A. Podosokorskaya, and I.V. Kublanov. Complete genome sequence of and proposal of *Thermofilum uzonense* sp. nov. a novel hyperthermophilic crenarchaeon and emended description of the genus *Thermofilum*. *Standards in Genomic Sciences*, 10:122, 2015.
- [147] B.J. Tully, J.B. Emerson, K. Andrade, J.J. Brocks, E.E. Allen, J.F. Banfield, and K.B. Heidelberg. *De Novo* sequences of *Haloquadratum walsbyi* from lake Tyrrell, Australia, reveal a variable genomic landscape. *Archaea*, 2015:875784, 2015.

- [148] Ľ. Valík, A. Medved'ová, M. Čížniar, and D. Liptáková. Evaluation of temperature effect on growth rate of *Lactobacillus rhamnosus* GG in milk using secondary models. *Chemical Papers*, 67:737–742, 2013.
- [149] L.V. Vasilyeva, M.V. Omelchenko, Y.Y. Berestovskaya, A.M. Lysenko, W.-R. Abraham, N. Dedysh, and G.A. Zavarzin. *Asticcacaulis benevestitus* sp. nov., a psychrotolerant, dimorphic, prosthecate bacterium from tundra wetland soil. *International Journal of Systematic and Evolutionary Microbiology*, 56:2083–2088, 2006.
- [150] A.D. Warth. Relationship between the heat resistance of spores and the optimum and maximum growth temperatures of *Bacillus* species. *Journal of bacteriology*, 134:699–705, 1978.
- [151] J.D. Watson and F.H.C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.
- [152] G.F. Weiller, G. Caraux, and N. Sylvester. The modal distribution of proteins isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics*, 4:943–949, 2004.
- [153] G. Wildgruber, M. Thomm, H. König, K. Ober, T. Ricchiuto, and K.O. Setter. *Methanoplanus limicola*, a plate-shaped methanogen representing a novel family, the methanoplanaceae. *Arch. Microbiol.*, 132:31–36, 1982.
- [154] R.A.D. Williams, K.E. Smith, S.G. Welch, and J. Micallef. *Thermus oshimai* sp. nov., isolated from hot springs in Portugal, Iceland, and the Azores, and comment on the concept of a limited geographical distribution of *Thermus* species. *International Journal of Systematic and Evolutionary Microbiology*, 46:403–408, 1996.
- [155] R.A.D. Williams, K.E. Smith, S.G. Welch, J. Micallef, and R.J. Sharp. DNA relatedness of *Thermus* strains, description of *Thermus brockianus* sp. nov., and proposal to reestablish *Thermus thermophilus* (Oshima and Imahori). *International Journal of Systematic Bacteriology*, 45:495–499, 1995.
- [156] J. Xu, X. Qiu, J. Dai, H. Cao, M. Yang, J. Zhang, and M. Xu. Isolation and characterization of a *Pseudomonas oleovorans* degrading the chloroacetamide herbicide acetochlor. *Biodegradation*, 17:219–225, 2006.
- [157] E. Yabuuchi, Y. Kosako, H. Oyaizu, I. Yano, H. Hotta, Y. Hashimoto, T. Ezaki, and M. Arakawa. Proposal of *Burkholderia* gen. nov. and transfer of seven species of the genus *Pseudomonas* homology group II to the new genus, with the type species *Burkholderia cepacia* (PALLERONI and HOLMES 1981) comb. nov. *Microbiol. Immunol.*, 36:1251–1275, 1992.
- [158] T. Zavřel. *Optimization of cultivation conditions for selected microalgae species focused on biomass and valuable substances production*. PhD thesis, Masaryk University, Brno, 2015.
- [159] W. Zhao, X. Zeng, and X. Xiao. *Thermococcus eurythermalis* sp. nov., a conditional piezophilic, hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-immersed chimney in the

- Guaymas Basin. *International Journal of Systematic and Evolutionary Microbiology*, 65:30–35, 2015.
- [160] X. Zhao, Z. Zhang, J. Yan, and J. Yu. GC content variability of eubacteria is governed by the pol III α subunit. *Biochemical and Biophysical Research Communications*, 356:20–25, 2007.
- [161] W. Zillig, K.O. Stetter, S. Wunder, W. Schulz, H. Priess, and I. Scholz. The *Sulfolobus*-“*Caldariella*” group: Taxonomy on the basis of the structure of DNA-dependent RNA polymerases. *Arch. Microbiol.*, 125:259–269, 1980.