Life History Traits and Genome Structure: Aerobiosis and G+C Content in Bacteria

Jean R. Lobry

Université Claude Bernard - Lyon I
Laboratoire de Biométrie, Biologie Évolutive
CNRS UMR 5558 - INRIA Helix project
43 Bd 11/11/1918
F-69622 VILLEURBANNE CEDEX, FRANCE
look@my.home.page.invalid,
http://pbil.univ-lyon1.fr/members/lobry/

Abstract. Evolution is a tinkerer not an engineer: the term exaptation was coined to signify that old structures, that could be not significant in terms of fitness, get re-used when environmental conditions changed. Here I show that the average protein composition of G+C rich bacteria were exapted to the switch from anaerobic to aerobic conditions. Because the proteome composition is under the strong control of directional mutation pressure, this is an example of exaptation at the molecular level for which the underlying mechanism is documented.

1 Introduction

During the last 20 years, genomic sequence data have been produced in a exponential way, with a doubling time close to 18 months, reminiscent of Moore's law in computer sciences (Fig. 1). We don't know whether this is just an anecdotical coincidence or evidence that the limiting factor for the production of genomic sequence data is related to computer performances, although the latter interpretation is my favorite given the perpetual struggle for disk space we are facing in my laboratory just to store primary data.

Whatever the underlying reason for this doubling time, we have a huge amount of data available and the problem is how to make sense from this. This paper is an example, admitly modest, of what is called data mining, or post-mortem data analysis, in which I have used previously published results to interpret them my own way. This paper is basically an attempt to make a connection between two previously published results that are summarized thereafter to provide background material.

1.1 Some Biological Terms

- Bacteria: this is a subset of living organisms on Earth. It is used here in its broad sense (*i.e.* Archae + Eubacteria) to designate small unicellular organisms without complex subcellular structures such as a nucleus.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 679–686, 2004.

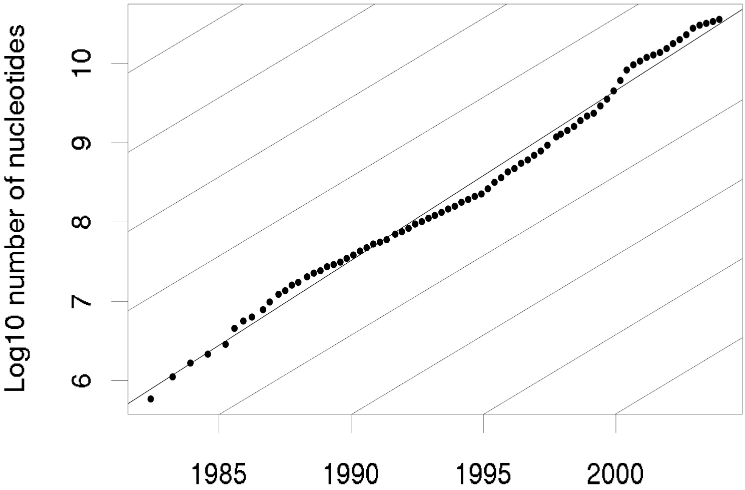

Fig. 1. The exponential growth of genomic sequence data mimics Moore's law. The source of data is the december 2003 release note (realnote.txt) from the EMBL database available at *http://www.ebi.ac.uk/*. External lines correspond to what would be expected with a doubling time of 18 months. The central line through points is the best least square fit, corresponding to a doubling time of 16.9 months.

– Aerobic bacteria: is used here to designate bacteria that can live only in presence of oxygen in their environment.
– Anaerobic bacteria: is used here to designate bacteria that can live only in absence of oxygen in their environment.
– Exaptation. Modern evolutionary theories are all based on Markov processes in which the future is influenced by the past only through the present state. To avoid using the term preadaptation, that would be misleading in this context as it may suggest some kind of knowledge of the future, Gould and Vrba have introduced the term exaptation [1]. This term is used to designate features of organisms that are non-adapted, but available for useful cooptation in descendants.

1.2 Genomic G+C Content and Aerobiosis

The G+C content is an example of global genomic structure that was used early in bacterial taxonomy, *i.e.* before the genomic era, because it was possible to estimate its value experimentally without knowing the sequence of a genome. The G+C content of bacterial chromosomes is the molar ratio of bases G and C over all bases, so that we could express this by the following lines of R implementation [2] of the S language:

```
> urn <- c("A", "C", "G", "T") # the bases in DNA
> dna <- sample(urn, size = 1000, replace = TRUE) # simulated DNA
> gc.content <- function(dna) {
    length(dna[dna == "G" | dna == "C"])/length(dna)
}
> gc.content(dna) # should be close to 0.5 in this case
[1] 0.467
```

Observed values range from 0.25 to 0.75 in bacteria and this was interpreted early as the result of differences in mutation rates between AT and GC pairs (see [3] and references therein). In the last 40 years, all the attempts to find an adaptative value for the G+C content of bacterial chromosome have failed, for instance there is no connection with the optimum growth temperature of bacteria [4] despite one may have expected from the extra hydrogen bond in GC pairs as compared to AT pairs.

Recently, Naya *et. al.* showed [5] that the G+C content is undoubtedly higher in aerobic bacteria than in anaerobic bacteria, linking for the first time a genome structure and a life history trait, and then raising the exciting possibility of a non-zero impact of the genomic G+C content on the cell fitness in bacteria. This kind of relationship between a genome structure and a life history trait is typically what makes sense for biologists because they are always looking for evidences of adaptation. Strickly speaking aerobiosis is not a life history trait *per se* but through its consequences because growth in aerobic conditions is much more efficient than in anaerobic conditions, allowing for smaller generation times.

1.3 Protein Metabolic Cost in Aerobic Conditions

Recently, Akashi and Gojobori have shown [7] that proteins produced in high amounts (*e.g.* ribosomal proteins) tend to avoid amino-acids that are expensive in terms of metabolic cost in aerobic conditions. This is an evidence that amino-acid composition of proteins is under the control of natural selection to enhance metabolic efficiency. On an other hand, the influence of the G+C content on the average amino-acid composition of proteins has been documented for a long time (see [6] and references therein): in G+C rich genomes, the encoded proteins tend to use amino-acid that are coded by G+C rich codons. Three groups of amino-acids can be defined to reflect their expected dependence on the G+C content [6]. A visual representation of Akashi and Gojobori data [7] taking into account these three groups of amino-acids is given in Fig. 2 that was generated with the following S code:

```
cost <- list(Ile = 32.3, Phe = 52.0, Lys = 30.3, Tyr = 50.0,
Asn = 14.7, Leu = 27.3, Met = 34.3, Asp = 12.7, Glu = 15.3,
Ser = 11.7,  Val = 23.3, Thr = 18.7, His = 38.3, Gln = 16.3,
Cys = 24.7,  Trp = 74.3, Arg = 27.3, Ala = 11.7, Pro = 20.3,
Gly = 11.7)gc.groups <- factor(rep(c(1, 2, 3), c(6, 10, 4)),
ordered = TRUE, label = c("low G+C", "mid G+C", "high G+C"))
stripchart(unlist(cost)~gc.groups, pch = 19, ylim = c(0.5,3.5),
xlim = c(0, max(unlist(cost))) )bx <- boxplot(unlist(cost)
~gc.groups, horizontal = TRUE, add = TRUE)
```

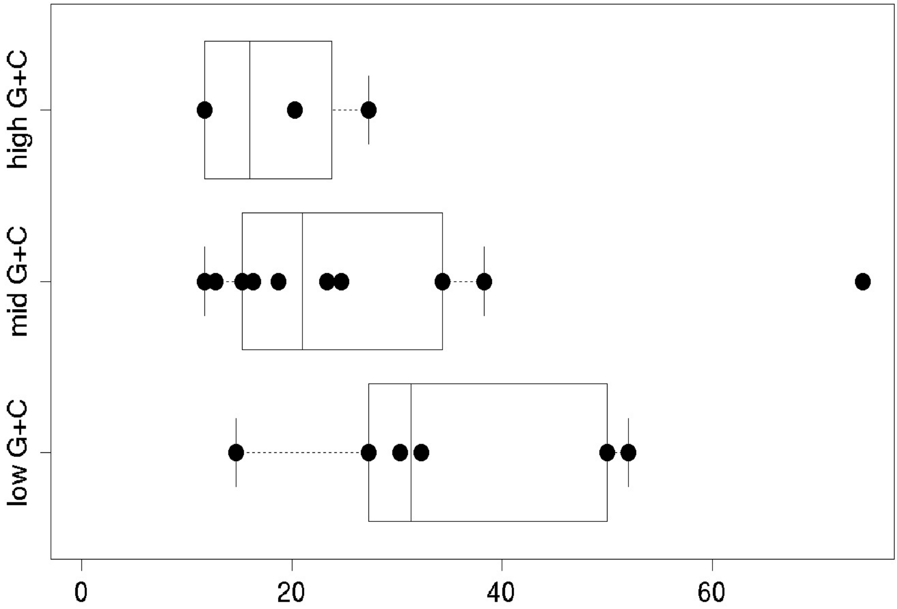

Fig. 2. Metabolic cost of the 20 amino-acids expressed in high-energy phosphate bond equivalent, ∼P, per amino-acid. Data are from table 1 in [7]. The box-and-whisker plot is a simple summary of data: the box represents the first quartile, the median and the last quartile ; the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

Fig. 2 shows that there is a trend for amino-acids that are favoured in high G+C genome to be less expensive in term of metabolic cost than those favoured low G+C genomes. It is therefore tempting to connect this with the result from Naya *et. al.* showing [5] that the G+C content is higher in aerobic bacteria. This is however not sufficient to conclude because we have to take into account the frequencies of amino-acid in proteins. For instance, the rigth outlier in the middle G+C group corresponds to Trp (*i.e.* tryptophan) which is known to be one of the rarest amino-acid in proteins (*cf* for instance table 1 in [8]).

2 Material and Methods

2.1 Source of Data

The G+C content in 225 anaerobic and 326 aerobic bacteria is from [5] and was download from `http://oeg.fcien.edu.uy/` and copied at: `http://pbil.univ-lyon1.fr/R/donnees/gc02.txt`. The amino-acids metabolic costs are from table 1 in [7]. The amino-acid frequencies in the proteins of 293 bacteria are from [9] and are available at: `ftp://pbil.univ-lyon1.fr/pub/datasets/JAG2003/`. This dataset, based on GenBank [10] release 130 including daily updates on the date of 13-JUL-2002, contains 97,095,873 codon counts.

2.2 Data Analyses

All analyses were done under R [2]. R is an Open Source implementation of the S language and similar to the commercial implementation S-Plus. S is both a general programming language and an extensible interactive environment for data analysis and graphics. See `http://www.r-project.org` for information on the project and CRAN (the Comprehensive R Archive Network) `http://cran.r-project.org` for available software and packages.

The model that predicts the frequency of a given amino-acid, aa, as function of the G+C content, θ, in absence of selective constraints is defined by:

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

with $\theta \in [0,1]$ and

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

This is a simple probabilistic model in which coding sequences are generated by random sampling from a DNA urn with a given G+C content. The numerator reflects the structure of the genetic code and the denominator is a correcting factor due to stop codons (see [6] for details).

To allow for the reproducibility of the results presented here, the R source code that was used to produce Fig. 3 is available at the URL: `http://pbil.univ- lyon1.fr/members/lobry/exapt/fig3.R`. If you don't have R at hand you can copy and paste this script in our RWeb interface at the following URL: `http://pbil.univ-lyon1.fr/Rweb/Rweb.general.html`.

3 Results and Discussion

3.1 Results

Results are summarized by Fig. 3 which is divided into two panels sharing as common x-axis the genomic G+C content, ranging from 0.25 to 0.75, as expected in bacteria.

The bottom panel of Fig.3 recalls Naya *et. al.* recent breakthrough [5]: the G+C content is higher in aerobic bacteria (on the right) than in anaerobic bacteria (on the left). A direct representation of data, with a small amount of noise

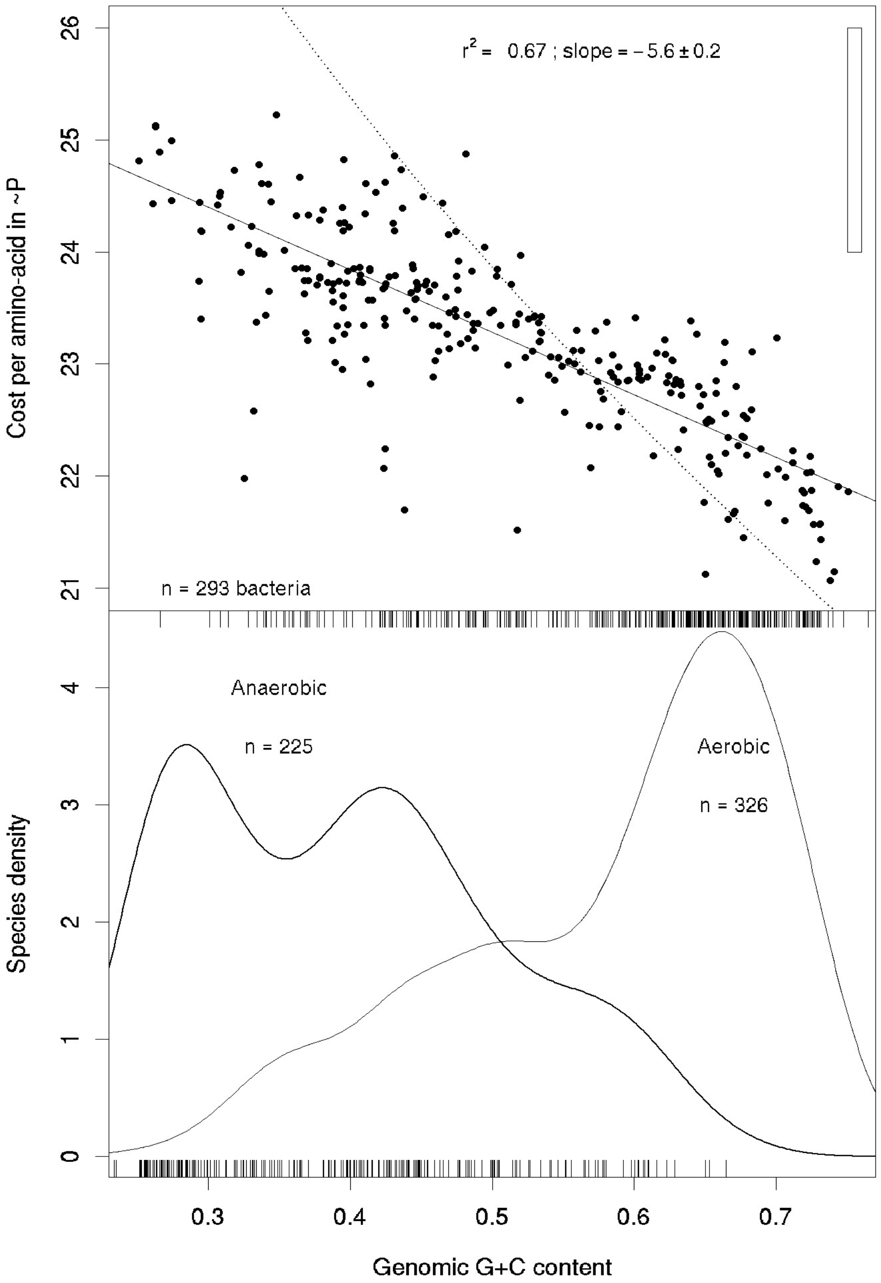

Fig. 3. Decrease of the average protein aerobic cost and distribution of anaerobic and aerobic bacteria with G+C content. See text for explanations.

added to break ties, is given on the top for aerobic species and at the bottom for anaerobic species. Although the two distributions are overlapping, there is clearly a trend for aerobic species to be G+C rich.

The top panel of Fig. 3 has a common y-scale expressed in aerobic metabolic cost (in high-energy phosphate bond equivalent, ∼P, per amino-acid) for the four following items:

1. The bar on the top right gives the range of observed selective effects [7] between proteins with a $\sim 10^5$-fold difference in terms of intracellular concentration. This bar is important to show a biologically relevant scale: a 2 ∼P per amino-acid difference in aerobic cost is enough to be selected in highly expressed genes. This bar gives also an idea of the within-species variability for the average protein aerobic cost.
2. The doted line represents what would be the average protein aerobic cost if protein composition was under the sole control of directional mutation pressure [6]. This model shows that under neutral conditions, if there were no selective constraints on the average amino-acid composition of protein, there would be an interest of being G+C rich under aerobic conditions because the cost decreases significantly (as compared to the the reference bar) when the G+C content increases. Note that the observed trend has a lower intensity, as expected, because the average protein composition is not completely free of selective constraints, so that the model is not realist.
3. The points represent the average (uniform protein weighting) aerobic cost for 293 bacteria, whose protein composition was deduced from a previously described dataset [9]. The actual average cost for a cell is expected to be lower because the uniform protein weighting is not realist. We should weight individual protein composition to take into account their intra-cellular concentrations in the cell, but this information is not available.
4. The line is the best least-squares fit: there is a significant decrease from low G+C to high G+C bacteria: from 24.7 to 21.9 ∼P per amino-acid. This 2.8 ∼P per amino-acid variation compares well the within-bacteria variation between highly and poorly expressed proteins [7] depicted by the reference bar.

3.2 Discussion

As noted by an anonymous referee of this paper, and I would like to take this opportunity to thanks him/her for valuable suggestions, I have assumed in the following discussion thatdirectional mutational pressure (responsible for generation of G+Ccontent) is free from selection. However, we can not exclude the possibility that themutational pressure is subjected in some way to selection (*e.g.* repairsystems may be selected to prefer some mutational defects than other leading tocomposition bias). I think this is unlikely, but even if this was true we would still have an example of exaptation at the molecular level. Features coopted as exaptations have two possible previous statuses. They may have been adaptations for another function, or they may have been non-adaptative features (*cf* section VI C in [1]).

It would be tempting to connect the top and the bottom of Fig. 3, assuming that no confounding factor is present, by a simple regular selective scenario: aerobic low cost amino acid are encoded by G+C rich codons so that the selection for low cost amino-acids at the proteome level has induced a G+C enrichment in coding sequences. This is, however, not defendable because in G+C rich bacteria the whole genome, including non-coding regions and synonymous positions, are also enriched in G+C content (*cf* [3] and references therein). The selective advantage results from the long-term effects of a directional mutation pressure. This is an example of exaptation at the molecular level: having a high G+C content is interesting under aerobic conditions, but this was unforeseeable before the oxygen concentration was enough on Earth.

References

1. Gould, S.J., Vrba, E.S.: Exaptation-A missing term in the science of form. Paleobiology **8** (1982) 4–15
2. Ihaka, R., Gentleman, R.: R: A Language for Data Analysis and Graphics. J. Comp. Graph. Stat. **3** (1996) 299–314
3. Lobry, J.R., Sueoka, N.: Asymmetric directional mutation pressures in bacteria. Genome Biology **3** (2002) 58.1–58.14
4. Galtier, N., Lobry, J.R.: Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. **44** (1997) 632–636
5. Naya, H., Romero, H., Zavala, A., Alvarez, B., Musto, H.: Aerobiosis increases the genomic guanine plus cytosine content (GC %) in prokaryotes. J. Mol. Evol. **55** (2002) 260–264
6. Lobry, J.R.: Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene **205** (1997) 309–316
7. Akashi, H., Gojobori, T.: Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **99** (2002) 3695–3700
8. Lobry, J.R., Gautier, C.: Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucl. Acids Res. **22** (1994) 3174–3180
9. Lobry, J.R., Chessel, D.: Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J. Appl. Genet. **44** (2003) 235–261
10. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: GenBank. Nucl. Acids Res. **30** (2002) 17–20