BAL Matthieu, GHEZAIEL Morad : Master 2 DLAD.

DNA SEQUENCE ASYMMETRY :

a comparative genomic study.

# **Introduction:**

Nowadays, international consortiums and frequently updated genome databases allows us to study various organisms from all reigns. Giving public or personal bioinformatic tools, complete genomes can be studied and compared to others. In this study, we will compare datas from three organisms : the prokaryote responsible for the anthrax , *Bacilius anthracis*, a cyanobacteria named *Oscilatoria acuminata* and the plant model *Arabidopsis thaliana*. On the first hand we will compare GC-skews profils for each species. On a second hand we will study specificity in dinucleotides distribution by comparing empiric and random generated datas. Intra and inter specific comparaisons will be realized and inferred with use of statistical test. In this comparative study, we will try to support endosymbiotic theory by joining GC-skew profiling and genomic signature datas.

# **Context:**

**DNA** structure is shaped to minimize molecular entropy and waste of information. Chargaff parity rules emitted before and after Watson and Crick discovery are evidence for an intrinsic stability in DNA composition. Studying deviation in such composition is a well known tool to identify origin *ori* and termination *ter* sites in a given sequence [1]. These biases can also be used in comparative studies. They are mutational events leading to sequence asymmetry and rupture of the Chargaff second parity rule (PR2) stating the existence of equivalent nucleotide distribution in DNA primary structure, as in its secondary structure following these rules (A=T and G=C). These events tend to occur during biological processes involving DNA as transcription and replication[1][2]. During replication, 2 replication forks start from a single or multiple origin sites and slide bidirectionally along the sequence. Replication stops when forks encounter theirselvse in a region containing the termination site. In double strand DNA, early strand is processed in the 5‘->3’ direction. Then the late strand is processed using Okasaki fragment, this, in order to follow the 5’->3’ rule. Even if replication induced bias are generally explained by differential repair activity between strands (polymerases access), it seems that helicase activity (concomitant to polymerase one) is responsible for mutational event. Adding the fact that thermodynamic entropy can lead to deamination process such as cytosine to thymidine conversion and we understand that nucleotide distribution is affected in these genomic areas. In that way, computing bias from different species may permit us to look for similar patterns at the organism level (genomic duplication) or between species (endosymbiont theory).

# **Material and methods :**

**Genome** assemblies from Sterne 34F2 strain of *Bacillius anthracis,* *Oscilatoria acuminata* and *Arabidopsis thaliana* were used in nucleotide distribution study. Sequences analysis were realized by the mean of a personal python script. FASTAS were parsed in order to pull genomic sequences. Then, reverse complementary strands wer determined. For the first pipeline, we computed GC-content, nucleotide distribution and GC-skew profiles (using the following formula : GC-skew = G-C/G+C). Preliminary tests were done to find the best window size that had to be used for GC-skew calculation. We find that defining a sequence window size between 0.2% and 2% long allowed us to easily identify *ter* et *ori* sites from *b.anthracis* without the use of signal transformation like Fourier transformation (final choice =0.2%). To check GC-skew sign evolution along the sequence we then computed cumulative graphs of GC-skew. Measures were made for both strands. Then, inter and intra specific statistical tests were done to compare GC-skew profiles (n=1/0.002=500). For this, Shapiro Wilk (normal distribution) and Bartlett (variance homogeneity) tests were realized before the 2 sample one way ANOVA. If p-values were <0.01 for each, ANOVA was realized, else, Kruskal and Wallis test was done. For dinucleotide distribution study, 10+20+40+80 =150 random generated sequence (4 random seeds) of 200 to 300 bp were used to calculate ratios between random and biological frequencies (observed frequency). Each 150 long sets were generated using the observed nucleotide frequency and the mean was computed. Then, we have proceeded to a Kruskal and Wallis test (n=16) to compare genomic signature between sequences. The following packages were used in the script: Matplotlib, Numpy, Scipy and Pandas.

# **RESULts:**

### GC-skew analysis as a comparative genomic tool

**To gain insight** into intra and inter-species sequence asymmetry, GC-skews and cumulative skews were computed for each specie.

Measurements realized for *b.anthracis* circular chromosome (fig1) show that GC-skew and cumulative values are increased after the 2000000th nucleotide indicating the *ori* site position. Proportional decrease is observed at the position 5000000 indicating here the *ter* site. This observation correlate to previous studies stating that bacterias have a unique ori site in their chromosomes, which is located in GC-skew sign shifting genomic area. Interestingly, results from the *b.anthracis* pOX1 plasmid (fig2) appear similar to those from its circular chromosome indicating an hypothetical intra-specific similarity for GC-skew profil. However, GC-skew standard deviations measured in *b.anthracis* sequences are quietly different(annexes) and these differences are easily visible by comparing GC-skew curves smoothness. By looking to results (fig3,4,5), we can see that GC skew biases in *o.acuminata* genome are more frequent than in b.anthracis genome. Consistently with previous hypothesis, *o.acuminata* chromosome shows a quietly similar GC-skew profile to pOSCIL1 ones (intra-specific similarity). According to previous studies, difference observed for the second plasmid pOSCIL2 could be explained by a more recent acquisition in phylogenetical time (both plasmids are needed for *b.anthracis* virulence). As expected, results (fig2,4,5) show us that GC-skew are highly frequent in plasmids. This observation is consistent with the idea that high GC-skew bias are preferentially found in organisms with significant horizontal transfer ability (plasmid transfer within sexual pili) according to the presence of multiple *ori* sites in plasmids. Finally, measurements made for *a.thaliana* linear chromosomes (fig6 to 10) display variable GC-skew profiles but a still high GC-skew rate. To get more insight into all these profiles, absolute GC-skew standard deviation were computed for each sequence (fig13). As expected, standard deviations seem to be significantly lower in linear chromosomes. These informations prompted us to propose that linear chromosomes possess GC-skews of high frequency but low intensity, leading to compensation phenomenon between skews at the whole sequence level.

*Biological explanation of this could be understood in two points:*

1. Eukaryotic cells possess higher number of genes in their linear chromosomes leading to frequent skews because of the transcriptional activity
2. Discontinuity of replication mechanism in linear chromosomes is a consequence of the Chargaff second parity rule (PR2), leading to the presence of multiple *ori* sites in order to ensure an equivalent nucleotide distribution along the sequence.

Having selected an optimal window size of 0.2% long, 500 cumulative GC-skew values were pulled from each sequence. Giving these datas and in order to proceed to a more significant comparative study, we have choosen to apply a 2-samples one way ANOVA test between each sequence pair. Results are shown as the absolute value of the log p-value. Databars from each specie are colored according to the reign appartenance (red = prokaryota, blue = cyanobacteria, green = eukaryota). First, datas presented in fig14 show us that *b.anthracis* circular chromosome GC skew distribution correlate more with circular sequences (pOX1, pOSCIL1, *a.thaliana* cpDNA) than with linear ones. This result support the previous idea that GC-skew profiles tend to be similar with those from sequences presenting the same 3d conformation (circular, linear). However, results fig15 are challenging this hypothesis by showing that *b.anthracis* pOX1 is more similar to *a.thaliana* linear chromosomes and mtDNA than to its circular chromosome or pOSCIL1 and 2 from *o.acuminata.* Interestingly, *o.acuminata* chromosome GC-skew profile (fig16) correlate with *a.thaliana* first and fourth chromosome as well its mt and cpDNA. This observation could be considered as a proof of endosymbiotic theory, for which plants and algae common ancestor appeared with the endocytosis of a cyanobacteria by a primitive eukaryote cell. Results from *a.thaliana* linear chromosomes (fig17 to 21) are consistent with the previous hypothesis. Finally, GC-skew profil measured on *a.thaliana* mtDNA appears similar to *b.anthracis* pOX1 plasmid one. This final observation could be another proof for endosymbiotic theory, theory for which mitochondrias come from the endocytosis of a bacteria by a proto eukaryote.

Beyond the use of GC profiling for *ori* and *ter* sites identification, all theses results are evidence for a GC skew profile similarity between species, similarly structured sequences (circular vs linear) and inside a specie.

### Dinucleotide distribution analysis is relevant for mutational events study and organism characterization

**To gain more insight** into sequence asymmetry between species, ratio between random dinucleotide distributions and observed ones were computed for each biological sequence. Each random sequence were generated with the same mononucleotide distribution than its biological counterpart. As expected for *b.anthracis* chromosome, dinucleotide distribution are highly biased toward A and T (cf AT/GC content table) with low GC, TG and AA occurrence and high GG,TA and AC occurrences. Such observation could be explained by mutation events converting cytosine to guanine. In that scenario, selection pressure could be responsible for late G to A conversion, this, in order to maintain Chargaff’s PR2. However, low GC content is inconsistent with previous studies showing that only aerobic obligate organisms presents a high GC content (*b.anthracis* being an aerobic facultative organism)[3]. Interestingly, *b.anthracis* pOX1 plasmid shows a slightly different profil in DN distribution with high GC, CC and CA frequencies while the whole sequence is highly biased toward A and T (cf content table). An explanation for this is the presence of a pathogenicity island (PAI) located in the plasmid, according to previous studies. For *o.acuminata* chromosome and plasmids, GC and AT contents are approximately equivalent. DN distribution profils are equivalent between all 3 sequences, with low GG and AA frequencies. However, results from *a.thaliana* linear chromosomes are not consistent with AT and GC content (cf content table) with a high bias to word A and T challenged by high CG and GG ratios. According to previous studies, such inconsistences are caused by multiple phenomenons as a decreasing GC gradient along the 5’-3’ direction, exons and introns structures and content, and more recently recombination rates (genome biased GC conversion) [4]. Beside these difference, whole DN distribution profils appear to be similar between all 5 chromosomes, supporting our previous hypothesis of DNA composition similarity between sequence of a same organism. Consistently with current model of DNA replication in chloroplasts, CC, AC, CA, CG and GC are highly distributed in cpDNA sequence with less AA, TA, TG, AT, and TT. The main model of circular cpDNA replication states that frequent deamination occurs on adenines leading to adenine to hypoxanthine conversion. [6] According to the model, such intermediate is then converted in guanine during replication. Giving our results, we could imagine a scenario during which CC ,AC and CA act as buffers “regulating” GC/CG conversion. In that way, we could have following mutation events:

1. TT to CC then CC to AC = > GC
2. TA to CA => CG
3. TG => CG
4. AA to AC => GC

Finally, results from *a.thaliana* mtDNA are consistent with the largely observed fact that mtDNA are often biased toward A and T. However, this observation is challenged by DN distribution displaying a high distribution of AC, CC, CG and CA, and surprisingly a very low distribution of TT. This observation can be explained by a physico-chemical reason that have been describe by previous studies: In organisms exposed to sunlight, UV radiations induce thymine dimers formation. These dimers are converted into intermediate that may block the replication process. Hence, organisms with high GC contents have a selective advantage to avoid such DNA lesions.

In the final part of this results section we have proceeded to statistical tests between each sequence for their dinucleotide frequencies (n=16 dinucleotide ratios). Taken together, results for all sequences show a significant intra-specific similarity in dinucleotide distribution. However Results from *b.anthracis* indicate that its DN frequencies (chromosome + pOX1) correlate more with those from *a.thaliana* linear chromosomes and cpDNA than with *o.acuminata* ones. In another hand, results from *o.acuminata* indicate that its whole genome correlates more with *a.thaliana* mtDNA than b.anthracis. Even if intra-specific similarity in DN distrubtion is clearly shown in these results, they do not support the endosymbiotic theory.

# **CONCLUSION:**

**DNA** structure and composition provide various features that can be used in comparative studies. For all reigns, natural selection contributes to improve the fitness of a specie through various means. Particularly, important biological processes such as transcription and replication are responsible for mutation events. Such events occur in various genomic area and are influenced by multiples elements as physico-chemical conditions (UV induced mutation), thermodynamic constraints (Chargaff’s PR2) or selective pressure at the codon level. Obviously, such features can be compared between species in order to understand life organization (phylogeny) or origins (endosymbiosis theory). In that way, GC-skew profiling is an important tool to study transcription and replication induced biases. In this study, our results have shown that GC-skew variability along the sequence is higher in circular sequences than in linear ones. Interestingly, intraspecifc similarity in GC-skew profiles were found making it a relevant tool for comparative study. As expected by the endosymbiotic theory, GC-skews profils in *b.anthracis* are similar to *a.thaliana* linear chromosomes, cp and mtDNA. Additionally, results from *o.acuminata* displays a high correlation with *a.thaliana* cpDNA, making GC-skew profiling an important tool for comparative genomic. Finally, results from genome signature analysis brought us relevant results concerning dinucleotide distribution, leading us to emit mutation event hypothesis. Various studies talk about genome signature as a relevant genetic tool to understand mutation events or virulence potential. However, it seems more relevant to decipher organism characteristics than to use it in comparative genomics because of profiles variability. In that way, asymmetry and bias position study along the sequence seems to be a more relevant tool for comparative studies in term of robustness.

# REFERENCE:

1. Arakawa, Kazuharu and Masaru Tomita. “The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection” *Evolutionary bioinformatics online* vol. 3 159-68. 6 Sep. 2007
2. Fujimori, Shigeo et al. “GC-compositional strand bias around transcription start sites in plants and fungi” *BMC genomics* vol. 6 26. 28 Feb. 2005, doi:10.1186/1471-2164-6-26
3. Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. 2002;55:260–264. doi: 10.1007/s00239-002-2323-3.
4. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA. 1988;85:2653–2657. doi: 10.1073/pnas.85.8.2653.
5. Eugénie Pessia, Alexandra Popa, Sylvain Mousset, Clément Rezvoy, Laurent Duret, Gabriel A. B. Marais; Evidence for Widespread GC-biased Gene Conversion in Eukaryotes, Genome Biology and Evolution, Volume 4, Issue 7, 1 January 2012, Pages 675–682, <https://doi.org/10.1093/gbe/evs052>
6. Krishnan, Neeraja M and Basuthkar J Rao. “A comparative approach to elucidate chloroplast genome replication” *BMC genomics* vol. 10 237. 20 May. 2009, doi:10.1186/1471-2164-10-237

# ANNEXES:

# 

# 









