Gene Reports 8 (2017) 45–48

Contents lists available at ScienceDirect

Gene Reports

journal homepage: www.elsevier.com/locate/genrep

An overview on the DNA nucleotide compositions across kingdoms

Yabin Guo

Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China.

ARTICLE INFO

Keywords:
Nucleotide composition
GC content
Purine content
Thermophilicity
Genome

ABSTRACT

The DNA nucleotide compositions vary among species. This fascinating phenomenon has been studied for decades with some interesting questions remaining unclear. Recent years, thousands of genomes have been sequenced, but general evaluations on the nucleotide compositions across different phylogenetic groups are still absent. In this article, I analyzed 371 genomes from different kingdoms and provided an overview on DNA nucleotide compositions. A number of important topics were discussed, including GC content, DNA strand symmetricity, CDS purine content, codon usage, thermophilicity in prokaryotes and non-coding RNA genes. I also gave explanations to two long debated questions: 1) both genome GC content and CDS purine content are correlated with the thermophilicity in archaea, but not in bacteria; 2) the purine rich pattern of CDS in most species is mainly a consequence of coding requirement, but not mRNA interaction dynamics. This study provides valuable information and ideas for future investigations.

The DNA molecules in all organisms are composed of the same four nucleotides, A, T, G and C, while the ratios of the four nucleotides vary among species, which has been fascinating to people for nearly a century. In 1950s, Erwin Chargaff found that in DNA the number of G equals the number of C, and the number of A equals the number of T, which is known as the Chargaff's first rule (Chargaff et al., 1950, 1952). Now we know that it is correct in double strand DNA for the complement between purines and pyrimidines. In 1960s, Chargaff published his second rule (the second parity rule, PR2), which stated that in each DNA strand the A ratio roughly equals the T ratio, and the G ratio roughly equals the C ratio (Rudner et al., 1968). This rule has been proved largely true except in some small DNA molecules such as the mitochondrial (mt) DNAs. Beside the overall genomic nucleotide composition, the nucleotide composition of coding sequences (CDS) is also an important topic. Szybalski et al. (1966) and Smithies et al. (1981) found that DNA template strands have more pyrimidine nucleotides (i.e. RNAs are purine rich), which was later named the Szybalski's rule by Forsdyke (Dang et al., 1998; Lao and Forsdyke, 2000). Forsdyke claimed that Thermophiles strictly obey Szybalski'r rule and raised a *Politeness Hypothesis*, assuming that mRNA with higher purine content are “polite” to avoid undesired interactions, and mRNA of thermophiles need to be even more polite, because the entropy-driven reactions are more prone to happen under high temperature (Lao and Forsdyke, 2000). However, the results of further studies turned out to be paradoxical (Paz et al., 2004; Mahale et al., 2012). So far, the applicability of Szybalski's rule has not been proved.

Most of these studies were performed in the *pre-genomics era* and sometimes based on incomplete genomic data. During the recent ten years, thanks to the development of next generation sequencing technology, genomes of thousands of species were sequenced. Yet, there still lacks a global evaluation on the DNA nucleotide compositions across kingdoms (or domains). In this article, I analyzed 371 genomes (122 animals, 39 plants, 53 fungi, 32 protists, 25 archaea and 100 bacteria) and revealed a number of amazing facts and provided explanations for two long unsolved questions.

First, the GC contents of all the nuclear genomes were calculated (Fig. 1A, Table S1). The GC contents of animal genomes have the smallest diversity with an average of 40%, and more invertebrate genomes have lower GC contents than vertebrate genomes do. Most of the plant genomes analyzed here falls into two groups: the dicots (yellow fill) with lower GC contents and the grass family (Poaceae) monocots (blue fill) with higher GC contents (Kumari and Ware, 2013; Smarda et al., 2014). Banana (*M. acuminata*), the only non-Poaceae monocot analyzed (red fill) has a medium GC content between the two groups. There are three plant genomes have considerably higher GC contents. Actually, they are green and red algae instead of Embryophytes. Protists and prokaryotes are more complex phylogenetic groups and it is not surprising that their genome GC contents have higher diversity. Among all known genomic sequences, bacterium, *Anaeromyxobacter dehalogenans*, has the highest GC content (74.9%), while *Candidatus* Zinderia insecticola (a symbiont in spittlebugs) has the lowest GC content (13.5%), even lower than all known mitochon-

Abbreviations: CDS, coding sequences; APCC, average purine content of CDS; TI, Thermo index; mt, mitochondrion/mitochondrial

E-mail address: guoyb9@sysu.edu.cn.

http://dx.doi.org/10.1016/j.genrep.2017.05.003
Received 25 January 2017; Accepted 5 May 2017
Available online 09 May 2017

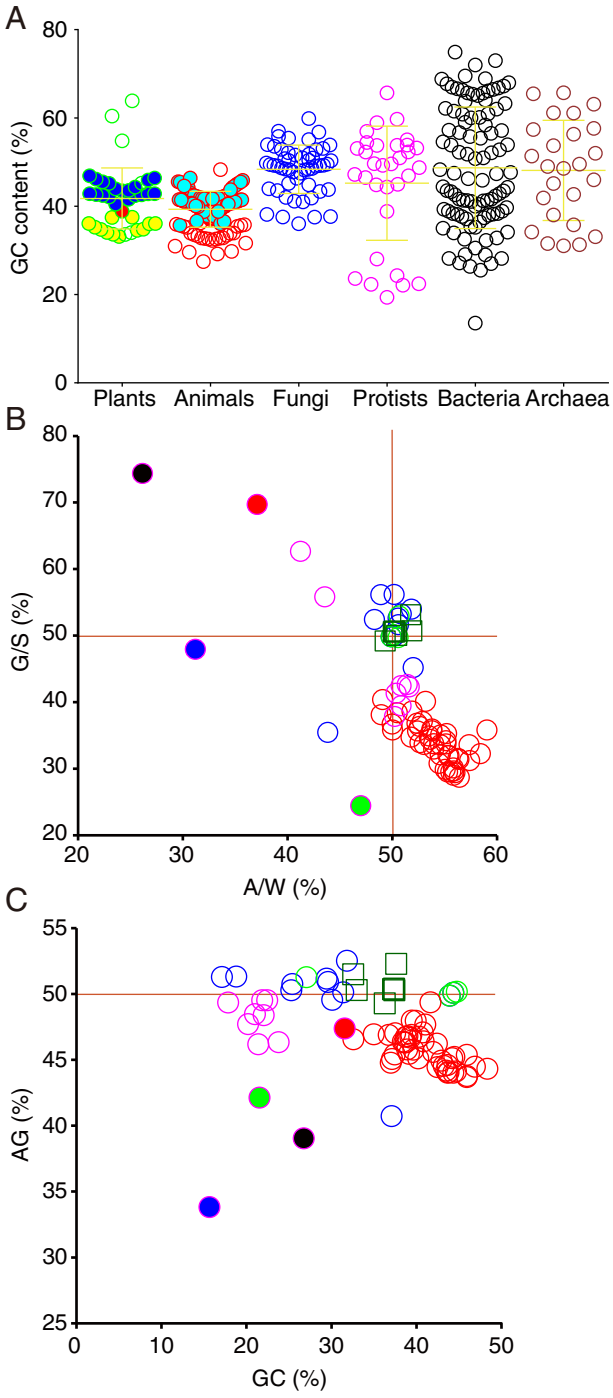

Fig. 1. Genome GC contents and mitochondrial nucleotide compositions. Each point is one species. A, genome GC contents across kingdoms (blue fill, plants of Poaceae; yellow fill, dicot plants; red fill, *Musa acuminata*; cyan fill, vertebrates); B, C, G/S-A/W (B) and AG-GC (C) plots for genomes of mitochondria and chloroplasts (red: vertebrates; magenta, invertebrates; green, plants; blue, fungi; dark green square, chloroplasts/plastids; blue fill, *Mnemiopsis leidyi*; green fill, *Atta cephalotes*; red fill, *Schistosoma mansoni*; black fill, *Onchocerca volvulus*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

drial genomes (Nishida, 2013). Archaea have relatively moderate DNA GC contents compared with bacteria, though many of them live in extreme environments. The genome of *Plasmodium falciparum* (one of the malaria parasite) has the lowest GC content (19%) in all eukaryotic genomes (Gardner et al., 2002).

Then, the Chargaff's second parity rule was evaluated. All large chromosomes are symmetric as expected (Table S1). Whereas, many mtDNAs have asymmetric strands as described previously (Francino and Ochman, 1997; Frank and Lobry, 1999). All animal mitochondrial genomes are small (10–20 kb, Fig. 1B, Table S1). The invertebrate mtDNAs have lower GC contents (21–32%) than those of vertebrates (32–49%). Although tunicate (*Ciona intestinalis*) is chordate and evolutionally much closer to vertebrates than to protostomes, its mtDNA nucleotide composition is more similar to those of protostomes (Fig. 1C). In a typical vertebrate mtDNA, one strand has more A and C, while the other stand has more G and T. The nucleotide compositions of invertebrate mtDNA have large diversity. The mtDNAs of leaf-cutting ant (*Atta cephalotes*, green fill), comb jelly (*Mnemiopsis leidyi*, blue fill), blood fluke (*Schistosoma mansoni*, red fill) and river blindness parasite (*Onchocerca volvulus*, black fill) are scattered far away from the cluster, matching their special places in evolution (Fig. 1B, C). The plant mtDNAs, chloroplast DNAs and fungal mtDNAs are usually larger and have more symmetric strands, distributing around point (50, 50) in the G/S-A/W plot (Fig. 1B).

To indicate how far a DNA strand is from symmetry more delicately, an index *Asy* (asymmetricity) is introduced (see Methods section). Briefly, *Asy* is the Euclidean distance between the point of a given DNA strand and the point (50, 50) on the G/S-A/W plot. Consisting with previously reported, the small mtDNAs have higher *Asy* values, but there is no correlation between *Asy* and mtDNA size (Fig. S1A). Among all the mtDNA analyzed, *O. volvulus* has the most asymmetric strands. Notably, the mtDNA of *Puccinia graminis* (the fungal pathogen of black rust in wheat) has a pretty high *Asy*, though its size is many times larger than typical animal mtDNAs (magenta fill in Fig. S1A, Table S1).

Small perturbations are found in the *Asy* values of prokaryotic and protist chromosomes, which is not surprising for their small sizes. And protist chromosomes show larger diversity in *Asy* than prokaryotic chromosomes do when their sizes are similar. One third of the Leishmania chromosomes have considerably high *Asy* values (Fig. S1B). Moreover, the G ratio correlates well with the A ratio in Leishmania chromosomal sequences, indicating there are heavy and light strands (Fig. S1C).

Statistically, larger chromosomes tend to have more symmetric strands than smaller ones do, while larger chromosomes also tend to have more asymmetric local regions. Indeed, the unassembled scaffolds/contigs of animal genomes show a substantially scattered pattern on the G/S-A/W plot (Fig. S2A). For example, in a 1.86 Mb scaffold of the kangaroo rat (*Dipodomys. ordii*) genome, the number of A is 8 times of that of T (*Asy* = 40.5). These unassembled scaffolds/contigs usually are highly repeated and many of them contain satellite DNA located in centromeres and telomeres. It is known that satellite DNA comprises more than half of the *D. ordii* genome (Mazrimas and Hatch, 1972). Besides *D. ordii*, large contigs with asymmetric strands can also be found in dog (*C. familiaris*) and collared flycatcher (*Ficedula albicollis*) genomes (Fig. S2B). Although many plant genomes contain large fractions of repetitive regions, similar highly asymmetric regions were not found in plant genomes (Fig. S2B).

Symmetry (high entropy) is more stable than asymmetry (low entropy). Inversions, translocations and transpositions make DNA strands more and more symmetric as time goes by (Albrecht-Buehler, 2006). Asymmetric strands in chromosomal regions (or small chromosomes) may be maintained by special replication mechanisms (e.g. for mtDNAs) or evolutionary benefits (e.g. for satellite DNAs), but it is impossible for a large chromosome to maintain asymmetric strands due to the high energy barrier. Actually, it is not surprising that Chargaff's second rule is correct, and it would be really surprising if it is not correct.

To test Szybalski's rule, I calculated the CDS of all the genomes mentioned above (Table S2). The GC contents of CDS correlate well with the GC contents of genomes, especially in prokaryotes, because CDS comprise most of the prokaryotic genomes, while in eukaryotes, the GC contents of CDS usually are greater than those of genomes (Fig. S3A). G/S-A/W plot shows that animal, plant and fungal CDS distribute in three different areas, while protist and prokaryotic CDS show large diversities (Fig. 2A). Similar to the previous reports in prokaryotes (Lao