Characteristics of oligonucleotide frequencies across genomes: Conservation versus variation, strand symmetry, and evolutionary implications

Shang-Hong Zhang*, Ya-Zhi Huang

The Key Laboratory of Gene Engineering of Ministry of Education, and Biotechnology Research Center, Sun Yat-sen University, Guangzhou 510275, China

Abstract

One of the objectives of evolutionary genomics is to reveal the genetic information contained in the primordial genome (called the primary genetic information in this paper, with the primordial genome defined here as the most primitive nucleic acid genome for earth’s life) by searching for primitive traits or relics remained in modern genomes. As the shorter a sequence is, the less probable it would be modified during genome evolution. For that reason, some characteristics of very short nucleotide sequences would have considerable chances to persist during billions of years of evolution. Consequently, conservation of certain genomic features of mononucleotides, dinucleotides, and higher-order oligonucleotides across various genomes may exist; some, if not all, of these features would be relics of the primary genetic information. Based on this assumption, we analyzed the pattern of frequencies of mononucleotides, dinucleotides, and higher-order oligonucleotides of the whole-genome sequences from 458 species (including archaea, bacteria, and eukaryotes). Also, we studied the phenomenon of strand symmetry in these genomes. The results show that the conservation of frequencies of some dinucleotides and higher-order oligonucleotides across genomes does exist, and that strand symmetry is a ubiquitous and explicit phenomenon that may contribute to frequency conservation. We propose a new hypothesis for the origin of strand symmetry and frequency conservation as well as for the constitution of early genomes. We conclude that the phenomena of strand symmetry and the pattern of frequency conservation would be original features of the primary genetic information.

Keywords: Compositional analysis; Whole-genome sequences; Primary genetic information; Frequency conservation; Strand symmetry

* Corresponding author. Tel.: 86-20-84110316, 86-20-84035425; E-mail: lsszsh@mail.sysu.edu.cn

1. Introduction

In the course of billions of years of evolution, organic genomes have undergone enormous changes. Nevertheless, some relics of the primordial genome may remain in modern genomes. Finding these relics is of great significance for the study of the origin and evolution of genomes.

What traits at the genomic level may be regarded as the relics? Candidates may include certain characteristics of very short sequences in modern genomes. For a sequence of DNA (or RNA) in a genome, the shorter it is, the less probable it would be modified during genome evolution. For example, the shortest possible sequences with more than one nucleotide, the dinucleotides, would have in general considerable chances to be intact (no mutations except being duplicated or deleted completely), provided that the rates of nucleotide substitution in cellular genomes are at the order around 10^{-9} substitutions per site per year (Li, 1997). For that reason, original features of the primordial genome, if still exist, would more probably be found in the characteristics of very short nucleotide sequences such as mononucleotides, dinucleotides, and higher-order oligonucleotides. One of such characteristics to be concerned about would be the pattern of genomic occurrence frequencies of these very short nucleotide sequences (whole-genome or large-scale duplications involve short and long sequences, but short nucleotide frequencies would generally be less, if any, influenced by these duplications if genomes are with good compositional homogeneity; the same would be true of large-scale deletions, though the specific loss of genome material in the evolution of some genetic systems such as organelle genomes would be somewhat different). If a considerable proportion (the number of copies) of a very short sequence has not been modified by nucleotide substitutions, insertions or deletions in evolving genomes, and/or if the variation of its occurrence frequencies has been limited by a certain mechanism (system) since the beginning or early stages of genome evolution, the pattern of genomic frequencies of the sequence would be conserved throughout the time, i.e., no significant changes since the primordial genome formed. From this point forward, our philosophy suggests that the conservation of the genomic pattern of the frequencies of mononucleotides, dinucleotides, and higher-order oligonucleotides across various genomes, if it exists, would probably be a relic of the primordial genome.

As for the frequencies of a mononucleotide across genomes, it has long been known that they vary among species, especially in prokaryotes (Sueoka, 1962). The frequencies of mononucleotides are usually described as GC content (percentage of G + C). The genomic GC content of prokaryotes may vary from less than 20% to over 70% (see below). The origin and evolution of genomic GC content is a fundamental problem in the study of genome evolution and a controversial issue (a subject to be discussed in separate papers). In spite of the variation of the GC content, it does not preclude the possibility of the conservation of the frequencies of some, if not all, dinucleotides and higher-order oligonucleotides across genomes. Many researches were done in the field of dinucleotide frequencies even when sequence data were limited (e.g., Nussinov, 1980; 1981; 1984), revealing hierarchies in the frequencies (preferences) of different dinucleotides in natural nucleic acid sequences. With more sequences available, one of the most studied aspects in the field of very short nucleotide frequencies is the characteristics of dinucleotide and higher-order oligonucleotide relative abundances, which access contrasts between the observed frequencies of short nucleotides and those expected from the frequencies of the components (Karlin and Burge, 1995; Karlin et al., 1997). The profiles of relative abundances of dinucleotides and higher-order oligonucleotides in genomic sequences are rather species-specific or taxon-specific (Karlin et al., 1994; Karlin et al., 1997; van Passel et al., 2006; Bohlin et al., 2008). The set of all dinucleotide relative abundance values is even regarded as a genomic signature (Karlin and Burge, 1995), a concept that may be extended to higher-order oligonucleotides (Deschavanne et al., 1999; Wang et al., 2005) or key combinations of oligonucleotide frequencies (Abe et al., 2003). This characteristic seems in contradiction with our assumption on the conservation of the pattern of frequencies. However, as assumed above, what we need for the purpose of our study is the occurrence frequencies, which are generally not congruent with the relative abundances (Burge et al., 1992). Moreover, instead of considering the frequencies of all dinucleotides or higher-order oligonucleotides in a genome as a whole, they should be analyzed one by one. Therefore, it is of interest to ascertain if the conservation in terms of occurrence frequencies of dinucleotides or higher-order oligonucleotides exists across genomes and to see if there is a mechanism for maintaining frequency conservation, or to determine to what extent their frequencies vary among species.

Another aspect of the pattern of frequencies of mononucleotides, dinucleotides, and higher-order oligonucleotides is the phenomenon of strand symmetry, which reflects the similarities of the frequencies of nucleotides and oligonucleotides to those of their respective reverse complements within single strands of genomic sequences. Traces of strand symmetry were first discovered by Chargaff and co-workers in the late 1960s (Karkas et al., 1968; Rudner et al., 1968). Therefore, strand symmetry is also called the second parity rule (Sueoka, 1995; Bell and Forsdyke, 1999) after the famous Chargaff's first parity rule (%A = %T and %C = %G in duplex DNA; Chargaff, 1951). The first parity rule has been fully explained by the Watson-Crick model for duplex DNA (Watson and Crick, 1953). In contrast, although considerable researches have been done on the ubiquitous phenomenon of strand symmetry, it is not fully recognized (Baisnée et al., 2002). Also, the issue on its origin and biological significance is controversial

(Baisnée et al., 2002; Forsdyke and Bell, 2004; Chen and Zhao, 2005; Albrecht-Buehler, 2006; Okamura et al., 2007). It is worth to note that in previous work, many analyses of strand symmetry focused on comparisons between the frequency of an oligonucleotide and that of its reverse complement, but a systematic survey (comparison) of the frequencies of any two oligonucleotides of the same order in genomic sequences, which is necessary for fully revealing the pattern of strand symmetry, has not been reported.

Based on the above assumption and reassessment, comparative analysis of the characteristics of frequencies of mononucleotides, dinucleotides, and higher-order oligonucleotides in the genomes of various organisms, including a systematic study of the pattern of strand symmetry, might provide insights into the features of the primordial genome as well as the primary genetic information it contained. With the development of genomics, more and more whole-genome sequences are now available, providing opportunities for the analysis of relics of the primordial genome at the genomic level. In this paper we analyzed the pattern of frequencies of mononucleotides, dinucleotides, and higher-order oligonucleotides of the whole-genome sequences from 458 species (including archaea, bacteria, and eukaryotes). We studied also the phenomenon of strand symmetry in these genomes with systematic frequency comparisons. The results show that the conservation of frequencies of some dinucleotides and higher-order oligonucleotides across genomes does exist, and that strand symmetry is a ubiquitous and explicit phenomenon contributing to this conservation. We conclude that the pattern of frequency conservation on one hand, and the pattern of strand symmetry on the other, would be original features of the primary genetic information.

2. Materials and methods

2.1. Whole-genome sequences

We downloaded the whole-genome sequence of every species of archaea and bacteria that was available as of November 2007 from the NCBI (ftp://ftp.ncbi.nih.gov/ genomes/). For the species that have two or more strains or subspecies whose genomes have been sequenced, only one was taken randomly from each of them (our analysis shows that the choice of samples does not influence the validity of the results, data not shown; we took only one sample from each species because we studied inter-specific pattern and because there was only one whole-genome sequence per species for most of the species analyzed). In total, 45 complete genomes of archaea and 395 complete genomes of bacteria were analyzed in the study (only their genus names and RefSeq accession numbers are presented, A denotes an archaeon): *Acidiphilium* (NC_009484.1), *Acidobacteria* (NC_008009.1), *Acidothermus* (NC_008578.1), *Acidovorax* (NC_ 08752.1, NC_008782.1), *Acinetobacter* (NC_009085.1, NC_005966.1), *Actinobacillus* (NC_009053.1, NC_009655.1), *Aeromonas* (NC_008570.1, NC_009348.1), *Aeropyrum*

(A) (NC_000854.2), *Agrobacterium* (NC_003062.2-3063.2), *Alcanivorax* (NC_008260.1), *Alkalilimnicola* (NC_008340.1), *Alkaliphilus* (NC_009633.1, NC_009922.1), *Anabaena* (NC_007413.1), *Anaeromyxobacter* (NC_007760.1, NC_009675.1), *Anaplasma* (NC_004842.2, NC_007797.1), *Aquifex* (NC_000918.1), *Archaeoglobus* (A) (NC_000917.1), *Arcobacter* (NC_009850.1), *Arthrobacter* (NC_008711.1, NC_008541.1), Aster yellows witches'-broom phytoplasma (NC_007716.1), *Azoarcus* (NC_008702.1, NC_006513.1), *Azorhizobium* (NC_009937.1), *Bacillus* (NC_009725.1, NC_003997.3, NC_004722.1, NC_006582.1, NC_002570.2, NC_006270.2, NC_009848.1, NC_000964.2, NC_005957.1), *Bacteroides* (NC_003228.3, NC_004663.1, NC_009614.1), *Bartonella* (NC_008783.1, NC_005956.1, NC_005955.1), *Baumannia* (NC_007984.1), *Bdellovibrio* (NC_005363.1), *Bifidobacterium* (NC_008618.1, NC_004307.2), *Bordetella* (NC_002927.3, NC_002928.3, NC_002929.2), *Borrelia* (NC_008277.1, NC_001318.1, NC_006156.1), *Bradyrhizobium* (NC_004463.1, NC_009485.1, NC_009445.1), *Brucella* (NC_006932.1-6933.1, NC_003317.1-3318.1, NC_009504.1-9505.1, NC_004310.3-4311.2), *Buchnera* (NC_004545.1), *Burkholderia* (NC_008390.1-8391.1-8392.1, NC_008060.1-8061.1-8062.1, NC_006348.1-6349.2, NC_006350.1-6351.1, NC_007509.1-7510.1-7511.1, NC_007650.1-7651.1, NC_009254.1-9255.1-9256.1, NC_007951.1-7952.1-7953.1), *Caldicellulosiruptor* (NC_009437.1), *Caldivirga* (A) (NC_009954.1), *Campylobacter* (NC_009802.1, NC_009715.1, NC_008599.1, NC_009714.1, NC_002163.1), *Candidatus Blochmannia* (NC_005061.1, NC_007292.1), *C. Carsonella* (NC_008512.1), *C. Methanoregula* (A) (NC_009712.1), *C. Pelagibacter* (NC_007205.1), *C. Protochlamydia* (NC_005861.1), *C. Ruthia* (NC_008610.1), *C. Vesicomyosocius* (NC_009465.1), *Carboxydothermus* (NC_007503.1), *Caulobacter* (NC_002696.2), *Chlamydia* (NC_002620.2, NC_000117.1), *Chlamydophila* (NC_004552.2, NC_003361.3, NC_007899.1, NC_002491.1), *Chlorobium* (NC_007514.1, NC_008639.1, NC_002932.3), *Chromobacterium* (NC_005085.1), *Chromohalobacter* (NC_007963.1), *Citrobacter* (NC_009792.1), *Clavibacter* (NC_009480.1), *Clostridium* (NC_003030.1, NC_009617.1, NC_009697.1, NC_009089.1, NC_009706.1, NC_008593.1, NC_003366.1, NC_004557.1, NC_009012.1), *Colwellia* (NC_003910.7), *Corynebacterium* (NC_002935.2, NC_004369.1, NC_003450.3, NC_007164.1), *Coxiella* (NC_002971.3), *Cytophaga* (NC_008255.1), *Dechloromonas* (NC_007298.1), *Dehalococcoides* (NC_002936.3, NC_009455.1, NC_007356.1), *Deinococcus* (NC_008025.1, NC_001263.1-1264.1), *Desulfitobacterium* (NC_007907.1), *Desulfococcus* (NC_009943.1), *Desulfotalea* (NC_006138.1), *Desulfotomaculum* (NC_009253.1), *Desulfovibrio* (NC_007519.1, NC_002937.3), *Dichelobacter* (NC_009446.1), *Dinoroseobacter* (NC_009952.1), *Ehrlichia* (NC_007354.1, NC_007799.1, NC_006831.1), *Enterobacter* (NC_009778.1, NC_009436.1), *Enterococcus* (NC_004668.1), *Erwinia* (NC_004547.2), *Erythrobacter* (NC_

007722.1), *Escherichia* (NC_000913.2), *Fervidobacterium* (NC_009718.1), *Flavobacterium* (NC_009441.1, NC_009613.1), *Francisella* (NC_006570.1), *Frankia* (NC_008278.1, NC_007777.1, NC_009921.1), *Fusobacterium* (NC_003454.1), *Geobacillus* (NC_006510.1, NC_009328.1), *Geobacter* (NC_007517.1, NC_002939.4, NC_009483.1), *Gloeobacter* (NC_005125.1), *Gluconobacter* (NC_006677.1), *Gramella* (NC_008571.1), *Granulibacter* (NC_008343.1), *Haemophilus* (NC_002940.2, NC_000907.1, NC_008309.1), *Hahella* (NC_007645.1), *Haloarcula* (A) (NC_006396.1), *Halobacterium* (A) (NC_002607.1), *Haloquadratum* (A) (NC_008212.1), *Halorhodospira* (NC_008789.1), *Helicobacter* (NC_008229.1, NC_004917.1, NC_000915.1), *Herminiimonas* (NC_009138.1), *Hyperthermus* (A) (NC_008818.1), *Hyphomonas* (NC_008358.1), *Idiomarina* (NC_006512.1), *Ignicoccus* (A) (NC_009776.1), *Jannaschia* (NC_007802.1), *Janthinobacterium* (NC_009659.1), *Kineococcus* (NC_009664.1), *Klebsiella* (NC_009648.1), *Lactobacillus* (NC_006814.2, NC_008497.1, NC_008526.1, NC_008054.1, NC_008530.1, NC_005362.1, NC_004567.1, NC_009513.1, NC_007576.1, NC_007929.1), *Lactococcus* (NC_002662.1), *Lawsonia* (NC_008011.1), *Legionella* (NC_006368.1), *Leifsonia* (NC_006087.1), *Leptospira* (NC_008510.1-8511.1, NC_004342.1-4343.1), *Leuconostoc* (NC_008531.1), *Listeria* (NC_003212.1, NC_003210.1, NC_008555.1), *Magnetococcus* (NC_008576.1), *Magnetospirillum* (NC_007626.1), *Mannheimia* (NC_006300.1), *Maricaulis* (NC_008347.1), *Marinobacter* (NC_008740.1), *Marinomonas* (NC_009654.1), *Mesoplasma* (NC_006055.1), *Mesorhizobium* (NC_002678.2, NC_008254.1), *Metallosphaera* (A) (NC_009440.1), *Methanobrevibacter* (A) (NC_009515.1), *Methanocaldococcus* (A) (NC_000909.1-1733.1), *Methanococcoides* (A) (NC_007955.1), *Methanococcus* (A) (NC_009635.1, NC_005791.1, NC_009634.1), *Methanocorpusculum* (A) (NC_008942.1), *Methanoculleus* (A) (NC_009051.1), *Methanopyrus* (A) (NC_003551.1), *Methanosaeta* (A) (NC_008553.1), *Methanosarcina* (A) (NC_003552.1, NC_007355.1, NC_003901.1), *Methanosphaera* (A) (NC_007681.1), *Methanospirillum* (A) (NC_007796.1), *Methanothermobacter* (A) (NC_000916.1), *Methylibium* (NC_008825.1), *Methylobacillus* (NC_007947.1), *Methylococcus* (NC_002977.6), *Moorella* (NC_007644.1), *Mycobacterium* (NC_002944.2, NC_002945.3, NC_009338.1, NC_002677.1, NC_008596.1, NC_009077.1, NC_008705.1, NC_008146.1, NC_000962.2, NC_008611.1, NC_008726.1), *Mycoplasma* (NC_009497.1, NC_007633.1, NC_004829.1, NC_000908.2, NC_006360.1, NC_006908.1, NC_005364.2, NC_004432.1, NC_000912.1, NC_002771.1, NC_007294.1), *Myxococcus* (NC_008095.1), *Nanoarchaeum* (A) (NC_005213.1), *Natronomonas* (A) (NC_007426.1), *Neisseria* (NC_002946.2, NC_003116.1), *Neorickettsia* (NC_007798.1), *Nitratiruptor* (NC_009662.1), *Nitrobacter* (NC_007964.1, NC_007406.1), *Nitrosococcus* (NC_007484.1), *Nitrosomonas* (NC_004757.1, NC_008344.1), *Nitrosospira* (NC_007614.1), *Nocardia* (NC_

006361.1), *Nocardioides* (NC_008699.1), *Nostoc* (NC_003272.1), *Novosphingobium* (NC_007794.1), *Oceanobacillus* (NC_004193.1), *Ochrobactrum* (NC_009667.1-9668.1), *Oenococcus* (NC_008528.1), Onion yellows phytoplasma (NC_005303.1), *Orientia* (NC_009488.1), *Parabacteroides* (NC_009615.1), *Paracoccus* (NC_008686.1-8687.1), *Parvibaculum* (NC_009719.1), *Pasteurella* (NC_002663.1), *Pediococcus* (NC_008525.1), *Pelobacter* (NC_007498.2, NC_008609.1), *Pelodictyon* (NC_007512.1), *Pelotomaculum* (NC_009454.1), *Photobacterium* (NC_006370.1-6371.1), *Photorhabdus* (NC_005126.1), *Picrophilus* (A) (NC_005877.1), *Polaromonas* (NC_008781.1, NC_007948.1), *Polynucleobacter* (NC_009379.1), *Porphyromonas* (NC_002950.2), *Prochlorococcus* (NC_005042.1), *Propionibacterium* (NC_006085.1), *Prosthecochloris* (NC_009337.1), *Pseudoalteromonas* (NC_008228.1, NC_007481.1), *Pseudomonas* (NC_002516.2, NC_008027.1, NC_004129.6, NC_009439.1, NC_002947.3, NC_009434.1, NC_004578.1), *Psychrobacter* (NC_007204.1, NC_007969.1, NC_009524.1), *Psychromonas* (NC_008709.1), *Pyrobaculum* (A) (NC_003364.1, NC_009376.1, NC_009073.1, NC_008701.1), *Pyrococcus* (A) (NC_000868.1, NC_003413.1, NC_000961.1), *Ralstonia* (NC_007347.1-7348.1, NC_007973.1-7974.1, NC_003295.1), *Rhizobium* (NC_007761.1, NC_008380.1), *Rhodobacter* (NC_007493.1-7494.1), *Rhodococcus* (NC_008268.1), *Rhodoferax* (NC_007908.1), *Rhodopirellula* (NC_005027.1), *Rhodopseudomonas* (NC_005296.1), *Rhodospirillum* (NC_007643.1), *Rickettsia* (NC_009881.1, NC_009883.1, NC_009879.1, NC_003103.1, NC_007109.1, NC_009900.1, NC_000963.1, NC_009882.1, NC_006142.1), *Roseiflexus* (NC_009767.1, NC_009523.1), *Roseobacter* (NC_008209.1), *Rubrobacter* (NC_008148.1), *Saccharophagus* (NC_007912.1), *Saccharopolyspora* (NC_009142.1), *Salinibacter* (NC_007677.1), *Salinispora* (NC_009953.1, NC_009380.1), *Salmonella* (NC_003198.1, NC_003197.1), *Serratia* (NC_009832.1), *Shewanella* (NC_008700.1, NC_009052.1, NC_007954.1, NC_008345.1, NC_009092.1, NC_004347.1, NC_009901.1, NC_009438.1, NC_009831.1, NC_008577.1, NC_008321.1, NC_008322.1, NC_008750.1), *Shigella* (NC_007613.1, NC_007606.1, NC_004337.1, NC_007384.1), *Silicibacter* (NC_003911.11, NC_008044.1), *Sinorhizobium* (NC_009636.1, NC_003047.1), *Sodalis* (NC_007712.1), *Solibacter* (NC_008536.1), *Sphingomonas* (NC_009511.1), *Sphingopyxis* (NC_008048.1), *Staphylococcus* (NC_002758.2, NC_004461.1, NC_007168.1, NC_007350.1), *Staphylothermus* (A) (NC_009033.1), *Streptococcus* (NC_004116.1, NC_009785.1, NC_004350.1, NC_003098.1, NC_004070.1, NC_009009.1, NC_009442.1, NC_006449.1), *Streptomyces* (NC_003155.3, NC_003888.3), *Sulfolobus* (A) (NC_007181.1, NC_002754.1, NC_003106.2), *Sulfurovum* (NC_009663.1), *Symbiobacterium* (NC_006177.1), *Synechococcus* (NC_006576.1, NC_008319.1, NC_007516.1, NC_007513.1, NC_007776.1, NC_007775.1, NC_009482.1, NC_009481.1, NC_005070.1,), *Synechocystis* (NC_000911.1), *Syntro-*

phobacter (NC_008554.1), *Syntrophomonas* (NC_008346.1), *Syntrophus* (NC_007759.1), *Thermoanaerobacter* (NC_003869.1), *Thermobifida* (NC_007333.1), *Thermococcus* (A) (NC_006624.1), *Thermofilum* (A) (NC_008698.1), *Thermoplasma* (A) (NC_002578.1, NC_002689.2), *Thermosipho* (NC_009616.1), *Thermosynechococcus* (NC_004113.1), *Thermotoga* (NC_000853.1, NC_009486.1), *Thermus* (NC_005835.1), *Thiobacillus* (NC_007404.1), *Thiomicrospira* (NC_007520.2, NC_007575.1), *Treponema* (NC_002967.9, NC_000919.1), *Trichodesmium* (NC_008312.1), *Tropheryma* (NC_004572.3), *Ureaplasma* (NC_002162.1), *Verminephrobacter* (NC_008786.1), *Vibrio* (NC_002505.1-2506.1, NC_006840.1-6841.1, NC_009783.1-9784.1, NC_004603.1-4605.1, NC_004459.2-4460.1), *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis* (NC_004344.2), *Wolbachia* endosymbiont of *Drosophila melanogaster* (NC_002978.6), *W.* endosymbiont of *Brugia malayi* (NC_006833.1), *Wolinella* (NC_005090.1), *Xanthobacter* (NC_009720.1), *Xanthomonas* (NC_003919.1, NC_003902.1, NC_006834.1), *Xylella* (NC_004556.1), *Yersinia* (NC_008800.1, NC_003143.1, NC_006155.1), *Zymomonas* (NC_006526.1), uncultured methanogenic archaeon (A) (NC_009464.1).

Along with the prokaryotic samples, 18 eukaryotic whole-genome sequences were used in the analysis. These included: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Candida glabrata* CBS138, *Cryptococcus neoformans* var. *neoformans* JEC21, *Debaryomyces hansenii* CBS767, *Drosophila melanogaster*, *Encephalitozoon cuniculi* GB-M1, *Eremothecium gossypii* ATCC 10895, *Guillardia theta* nucleomorph, *Homo sapiens*, *Kluyveromyces lactis NRRL* Y-1140, *Oryza sativa*, *Ostreococcus lucimarinus* CCE9901, *Pichia stipitis* CBS 6054, *Plasmodium falciparum* 3D7, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Yarrowia lipolytica* CLIB122. All the data were also downloaded from the NCBI.

In addition, we downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) the complete sequences of some mitochondrial genomes (64 samples), chloroplast genomes (42 samples), plasmids (prokaryotic, 96 samples; eukaryotic, 16 samples), viruses (64 samples), phages (64 samples), and viroids (32 samples). These data were analyzed to compare with the results obtained from the prokaryotic and eukaryotic genomes.

2.2. Calculations of occurrence frequencies of mono-, di-, tri-, tetra-, penta-, and hexanucleotides

We counted the number of occurrences of every mono-, di-, tri-, tetra-, penta-, and hexanucleotide in each cellular genome, and also that of every mono-, di-, and trinucleotide in each of the organelle genomes, virus genomes, and other sequences. The count was performed by moving the sliding window of 1, 2, 3, 4, 5, or 6 nt down the sequence one base at a time. Each chromosome or contig was analyzed separately, without concatenation. Counts were compiled for each species as well as for each eukaryotic chromosome. Occurrence frequencies (percentages) were calculated from these counts. The frequencies of items containing ambiguous bases were also calculated, but not taken into account because of their very small values. In the calculations of occurrence frequencies, overlapping di-, tri-, and higher-order oligonucleotides were counted. We used this approach not only because occurrence frequencies are usually calculated in this way, but also because results from previous related studies indicate that overlapping and non-overlapping data sets are similar and highly correlated (Rogerson, 1989), and that strand symmetry is true for overlapping and non-overlapping tuples (Prabhu, 1993). In the calculations, only one strand of each genome (the downloaded sequence) was analyzed. Although the choice of strands seems arbitrary, the characteristics of strand symmetry (Fickett et al., 1992; Prabhu, 1993; Qi and Cuticchia, 2001; Baisnée et al., 2002; Albrecht-Buehler, 2006) guarantee the validity of the results. In fact, there is little difference in terms of occurrence frequencies of mono-, di-, trinucleotides, and even higher-order oligonucleotides in analyzing one strand or another or both strands of a cellular genome (data not shown).

All the calculations were performed with computer programs written in C++.

2.3. Statistical analysis

To study whether the occurrence frequencies of short nucleotides are conserved across genomes, we employed the correlation/regression analysis to evaluate the correspondence between the observed values (data from each genome) and the expected values (hypothesized conserved values). For each mono-, di-, tri-, tetra-, penta-, or hexanucleotide, we analyzed the correlation between the observed counts and the expected counts in the genomes studied. The expected count of a nucleotide or oligonucleotide in a genome was obtained from the total count of all nucleotides or oligonucleotides of the same order in that genome multiplied by the mean frequency of that particular nucleotide or oligonucleotide in the genomes studied. We calculated the Pearson correlation coefficient (r), the slope and intercept of the best-fitted line for the observed counts vs. the expected counts. A correlation coefficient and a slope close to 1, and an intercept near the origin would indicate that the frequencies are well conserved across genomes. As several eukaryotic genomes are very large compared with prokaryotic genomes, the results of correlation/regression analysis may be biased. To reduce this bias, we used only the data of an individual chromosome from each of these eukaryotes (*A. thaliana*, chromosome 1; *C. elegans*, chromosome 5; *D. melanogaster*, chromosome 3L; *H. sapiens*, chromosome 21; *O. sativa*, chromosome 6). This approach is appropriate because the frequency data are generally very similar among chromosomes of a eukaryotic species (data not shown), implying that the frequency profile of an individual chromosome and that of the entire genome would be equivalent. In addition, we employed the *t*-test to ensure that the correlation/regression results (*r*, slope, and intercept), after standardization, are respectively and significantly different between the frequency-conserved short nucleotides and the frequency-varied ones.

In the study related to strand symmetry, we used also the correlation/regression analysis to assess the similarities/differences between the count profile of a mono-, di-, tri-, tetra-, penta-, or hexanucleotide and the count profile of another across genomes. The *t*-test was also employed to distinguish the correlation/regression results.

3. Results

3.1. Mononucleotide frequencies across genomes

The GC content varies from 16.6% to 74.9%, with a mean value of 49.1%, in the analyzed prokaryotic genomes. While in the analyzed eukaryotic genomes, it varies from 19.4% to 60.4%, with a mean value of 40.6%. Although the mononucleotide frequencies are variable across genomes, the phenomenon of mononucleotide strand symmetry is very obvious and universal in all the cellular genomes analyzed (see below).

3.2. Dinucleotide frequencies across genomes

The distribution pattern of the frequencies of 16 dinucleotides of 440 species of archaea and bacteria and 18 species of eukaryotes is shown in Fig. 1. It is clear that the frequency ranges of the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG (dinucleotides composed of one strong nucleotide and one weak nucleotide) are much narrower across genomes than those of other dinucleotides, indicating that the frequencies of AC, AG, CA, CT, GA, GT, TC, and TG would be more conserved than others. While the distributions of the frequencies of AA, AT, CC, CG, GC, GG, TA, and TT dinucleotides (consisting of two strong nucleotides or two weak nucleotides) are dispersed throughout their respective ranges, most of the genomic frequencies of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides are not far away from their own means. The frequencies of the dinucleotides CC, CG, GC, and GG in the analyzed eukaryotes are generally low due to relatively low GC content of most of the genomes, but they are very variable as well. This characteristic is also evident from the statistics such as the standard deviation, the coefficient of variation, the minimum and the maximum (with the mean close to the median, especially for the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG; Table 1). Furthermore, the correlation/regression analysis revealed that the observed counts and the expected counts are highly correlated and very similar across archaeal and bacterial genomes for the eight frequency-conserved dinucleotides mentioned above. Their correlation coefficients for the observed vs. expected counts are very close to 1 ($r > 0.96$, $P < 10^{-250}$), with also the slopes close to 1 and the intercepts relatively small. For the other eight dinucleotides (AA, AT, CC, CG, GC, GG, TA, and TT), the correlation coefficients are between 0.24 ($P < 10^{-6}$) and 0.88 ($P < 10^{-140}$), but the slopes are not close to 1 and the intercepts are relatively large (Table 1). Virtually the same conclusion could be drawn for dinucleotide frequencies across genomes of eukaryotes and across genomes of archaea, bacteria, and eukaryotes taken together (Table 1, the results for eukaryotic genomes only are somewhat biased and not very reliable because of sample limitation). Actually, given that the frequencies of a dinucleotide are conserved across genomes, so are those of its reverse complement, which is consistent with the phenomenon of strand symmetry. In addition, similar results were obtained for eukaryotic chromosomes when they were taken as units of analysis (data not shown).

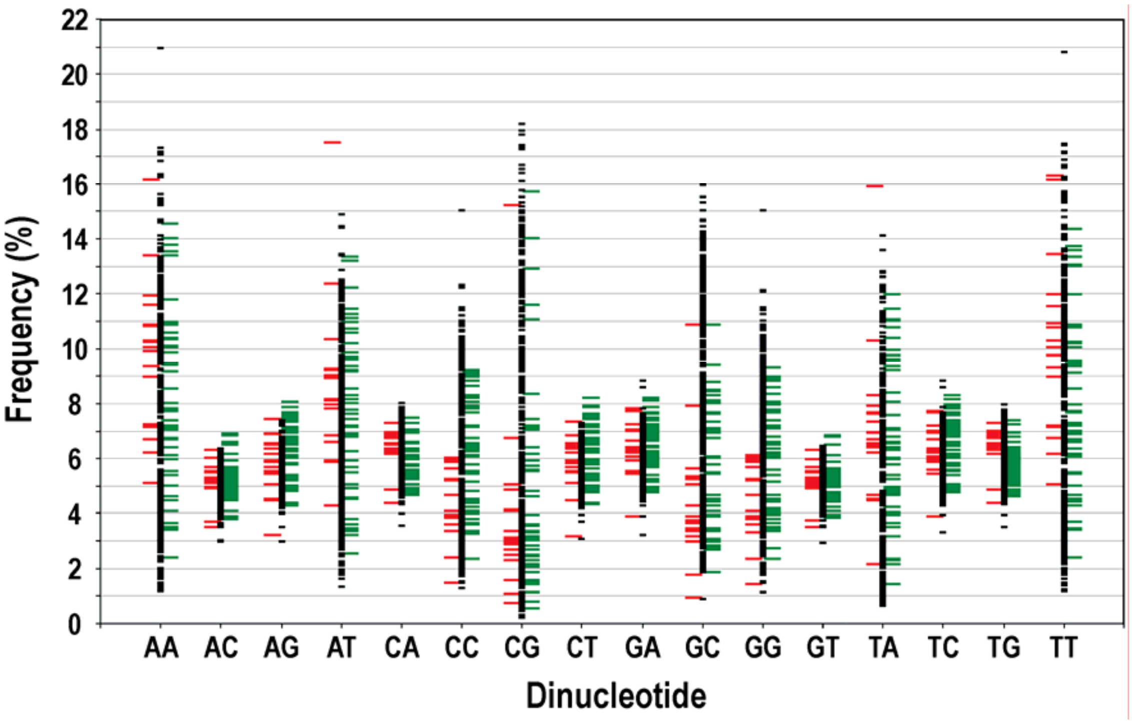

Fig. 1. Dinucleotide frequency distribution pattern of genomes of 395 species of bacteria (small black dash), 45 species of archaea (green dash), and 18 species of eukaryotes (red dash).

As our results show, there is a general correlation between the observed counts and the expected counts of a dinucleotide in the genomes studied, a correlation observed even for dinucleotides whose frequencies are not well conserved across genomes. This general correlation is mainly due to the usual trend that the observed counts of a dinucleotide increase with genome sizes. Therefore, what is important and interesting in our results is the observation that the observed counts and the expected counts of some dinucleotides are very highly correlated. This special correlation is due to frequency

Table 1 Statistical analysis of dinucleotide frequencies across genomes

Dinucleotide	Mean (%)	Min[a] (%)	Max[b] (%)	Median (%)	*s* [c]	CV[d] (%)	*r* [e]	Slope[f]	Intercept[f]
Archaea and bacteria, 440 genomes									
AA	7.96	1.19	20.94	8.01	4.04	50.78	0.42	0.38	126588.53
AC	4.97	2.98	6.87	5.02	0.59	11.94	0.98	1.08	-10578.81
AG	5.48	2.93	8.02	5.41	0.73	13.37	0.97	0.88	17631.85
AT	7.03	1.31	14.90	6.86	2.85	40.56	0.62	0.53	85635.99
CA	6.19	3.56	7.99	6.23	0.78	12.56	0.97	0.99	4954.80
CC	6.07	1.27	15.02	5.96	2.59	42.65	0.88	1.43	-67816.23
CG	6.83	0.21	18.20	5.85	4.49	65.78	0.84	1.95	-173920.89
CT	5.48	3.04	8.21	5.41	0.73	13.38	0.97	0.89	17415.89
GA	6.02	3.20	8.84	5.95	0.83	13.72	0.98	1.08	-11963.95
GC	7.50	0.88	15.95	7.06	3.62	48.23	0.88	1.65	-128709.89
GG	6.06	1.11	15.00	5.94	2.58	42.56	0.88	1.43	-67486.41
GT	4.97	2.91	6.85	5.00	0.59	11.92	0.98	1.08	-10692.07
TA	5.27	0.63	14.09	5.15	3.40	64.50	0.24	0.26	99698.09
TC	6.02	3.28	8.82	5.96	0.83	13.76	0.98	1.08	-12151.12
TG	6.18	3.48	7.95	6.22	0.78	12.55	0.97	0.99	4812.65
TT	7.95	1.16	20.79	7.96	4.05	50.88	0.42	0.39	126465.31
Eukaryotes, 18 genomes									
AA	10.11	5.08	16.16	10.12	3.09	30.54	0.90	1.07	-120599.98
AC	5.19	3.47	6.31	5.25	0.69	13.28	0.97	0.97	36279.43
AG	5.77	3.19	7.43	5.89	1.00	17.30	0.95	1.04	-42545.88
AT	8.66	4.25	17.49	8.53	2.87	33.10	0.84	1.04	-55997.24
CA	6.37	4.36	7.29	6.65	0.84	13.16	0.97	1.05	-42403.39
CC	4.30	1.44	6.03	3.99	1.33	30.99	0.87	1.05	-27430.17
CG	3.83	0.72	15.23	2.99	3.19	83.26	0.36	0.48	294077.27
CT	5.75	3.17	7.34	5.89	0.99	17.16	0.95	1.04	-41796.50
GA	6.32	3.88	7.79	6.27	0.90	14.22	0.96	0.91	77035.27
GC	4.47	0.90	10.86	3.82	2.24	50.19	0.73	0.86	95689.50
GG	4.30	1.42	6.09	3.97	1.36	31.53	0.86	1.04	-26153.15
GT	5.18	3.49	6.30	5.23	0.68	13.09	0.97	0.96	38410.51
TA	6.92	2.14	15.91	6.62	2.89	41.83	0.79	1.10	-96898.13
TC	6.31	3.87	7.70	6.27	0.88	13.98	0.96	0.91	77902.41
TG	6.37	4.37	7.27	6.64	0.84	13.14	0.97	1.05	-40392.33
TT	10.11	5.06	16.30	10.16	3.11	30.77	0.90	1.07	-111721.28
Archaea, bacteria, and eukaryotes, 458 genomes									
AA	8.05	1.19	20.94	8.09	4.03	50.07	0.86	1.12	-69497.90
AC	4.98	2.98	6.87	5.03	0.60	12.01	0.99	1.05	-5072.69
AG	5.49	2.93	8.02	5.43	0.75	13.58	0.98	1.03	-9343.29
AT	7.09	1.31	17.49	7.02	2.87	40.41	0.87	1.11	-50335.14
CA	6.20	3.56	7.99	6.27	0.78	12.58	0.98	1.04	-7236.40
CC	6.00	1.27	15.02	5.87	2.57	42.89	0.84	0.86	48452.64
CG	6.71	0.21	18.20	5.65	4.48	66.81	0.58	0.73	103168.92
CT	5.49	3.04	8.21	5.42	0.75	13.57	0.98	1.03	-9148.04
GA	6.03	3.20	8.84	6.00	0.83	13.76	0.98	1.02	-123.78
GC	7.38	0.88	15.95	6.91	3.62	49.05	0.73	0.77	91487.53
GG	5.99	1.11	15.00	5.87	2.56	42.80	0.84	0.86	48320.32
GT	4.97	2.91	6.85	5.02	0.60	11.98	0.99	1.05	-4946.80
TA	5.34	0.63	15.91	5.24	3.39	63.60	0.79	1.15	-57391.21
TC	6.04	3.28	8.82	5.98	0.83	13.79	0.98	1.02	383.82
TG	6.19	3.48	7.95	6.25	0.78	12.57	0.98	1.04	-7422.74
TT	8.04	1.16	20.79	8.01	4.03	50.17	0.86	1.12	-68659.13

[a] minimum; [b] maximum; [c] standard deviation; [d] coefficient of variation; [e] Pearson correlation coefficient for the

relationship between the observed counts and the expected counts (with all the *P* values < 0.001 except that of CG for eukaryotes only); [f] slope or intercept of the best-fitted line for the observed counts vs. the expected counts.

conservation across genomes of the dinucleotides concerned. The *t*-test can clearly distinguish the correlation/regression results of the frequency-conserved dinucleotides from those of the frequency-varied ones (for *r*, slope, and intercept, respectively, $P < 0.001$).

Last but not least, there is a specific pattern for the frequency conservation of dinucleotides across genomes. For example, in addition to the pattern that only dinucleotides composed of one strong nucleotide and one weak nucleotide are with well-conserved frequencies, among the eight frequency-conserved dinucleotides, AC and GT are more conserved across genomes than AG and CT. Moreover, there are differences among the means of the frequencies of some conserved dinucleotides (Fig. 1 and Table 1). For example, the mean of the frequencies of AC dinucleotide of the analyzed genomes is significantly different (smaller) from that of AG dinucleotide (*t*-test, $P < 0.001$).

3.3. Trinucleotide frequencies across genomes

The distribution pattern of the frequencies of 64 trinucleotides of 440 species of archaea and bacteria and 18 species of eukaryotes is shown in Fig. 2. Similar to the pattern of dinucleotides, the frequency ranges of some trinucleotides are much narrower across genomes than those of others. On the other hand, trinucleotide frequencies across genomes show a less conserved or a more variable pattern than dinucleotide frequencies. While the maximal variation of the genomic frequencies of a dinucleotide is limited to 87-fold (CG), it reaches 578-fold for TAT trinucleotide. Moreover, the mean of the coefficients of variation for dinucleotides is smaller than the same parameter for trinucleotides (see also Table 4). Statistics of trinucleotide frequencies of all the cellular genomes analyzed are presented in Table 2. The results of the correlation/regression analysis show that less than half of the 64 trinucleotides are with well-conserved frequencies across all prokaryotic and eukaryotic genomes analyzed (Table 2). Trinucleotides with the most conserved frequencies include: ATC, ATG, CAT, GAA, GAT, TCA, TGA, and TTC, containing one strong nucleotide and two weak nucleotides. All the conserved trinucleotides contain both strong and weak nucleotides, but no CG dinucleotide (the most variable dinucleotide). Also, when the frequency of a trinucleotide is conserved, so is the frequency of its reverse complement. The *t*-test can as well distinguish the correlation/regression results of the frequency-conserved trinucleotides from those of the frequency-varied ones (for *r*, slope, and intercept, respectively, $P < 0.001$).

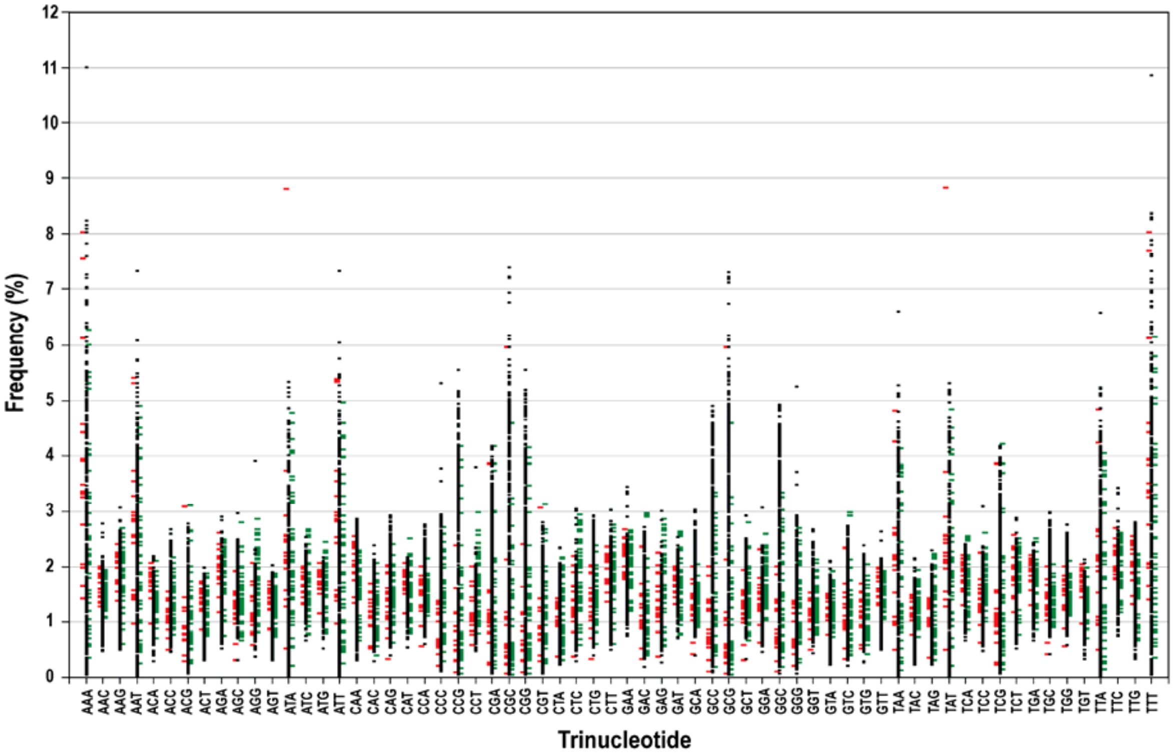

Fig. 2. Trinucleotide frequency distribution pattern of genomes of 395 species of bacteria, 45 species of archaea, and 18 species of eukaryotes. Symbols are as described in Fig. 1.

Table 2 Statistical analysis of trinucleotide frequencies across archaeal, bacterial, and eukaryotic genomes (458 genomes)

Trinucleotide	Mean (%)	Min[a] (%)	Max[b] (%)	Median (%)	*s*[c]	CV[d] (%)	*r*[e]	Slope[f]	Intercept[f]
AAA	2.79	0.04	10.99	2.58	1.97	70.55	0.78	1.21	-37817.07
AAC	1.47	0.49	2.78	1.52	0.38	25.57	0.95	1.04	-4466.16
AAG	1.70	0.50	3.06	1.78	0.53	30.95	0.93	1.01	-4921.79
AAT	2.08	0.03	7.32	1.88	1.40	67.38	0.80	1.15	-22292.41
ACA	1.25	0.29	2.18	1.28	0.34	27.37	0.94	1.36	-17535.67
ACC	1.40	0.46	2.66	1.42	0.42	30.32	0.90	0.87	9132.25
ACG	1.23	0.06	3.10	1.19	0.59	48.01	0.72	0.80	13570.92
ACT	1.10	0.31	1.97	1.14	0.40	36.76	0.92	1.20	-10240.15
AGA	1.43	0.53	2.89	1.36	0.49	34.26	0.92	1.21	-14535.54
AGC	1.64	0.32	2.94	1.68	0.40	24.68	0.91	0.82	12494.84
AGG	1.33	0.38	3.89	1.30	0.41	30.98	0.93	0.95	3005.62
AGT	1.10	0.31	2.02	1.14	0.40	36.72	0.92	1.20	-10308.15
ATA	1.69	0.02	8.81	1.55	1.26	74.41	0.68	1.31	-28049.17
ATC	1.75	0.67	2.66	1.79	0.29	16.74	0.97	0.90	6040.31
ATG	1.57	0.52	2.43	1.62	0.32	20.33	0.97	1.09	-6308.27
ATT	2.08	0.03	7.32	1.88	1.41	67.47	0.80	1.14	-22017.83
CAA	1.76	0.31	2.85	1.88	0.58	32.87	0.92	1.07	-8127.42
CAC	1.26	0.29	2.37	1.27	0.41	32.95	0.93	1.08	-433.17
CAG	1.61	0.34	2.92	1.61	0.54	33.27	0.90	0.94	7416.13
CAT	1.57	0.53	2.50	1.62	0.32	20.32	0.97	1.09	-6091.80
CCA	1.60	0.57	2.74	1.60	0.43	26.54	0.94	0.98	3716.76
CCC	1.23	0.10	5.30	1.13	0.71	57.89	0.78	0.87	9025.91
CCG	1.84	0.05	5.53	1.53	1.39	75.56	0.52	0.67	32818.18
CCT	1.33	0.48	3.78	1.29	0.41	30.93	0.93	0.95	2892.00

Table 2 (Continued)

CGA	1.59	0.06	4.16	1.37	0.98	61.53	0.65	0.81	19445.46
CGC	2.05	0.03	7.38	1.57	1.68	81.99	0.47	0.69	37369.25
CGG	1.84	0.04	5.52	1.53	1.39	75.50	0.52	0.67	32801.66
CGT	1.23	0.07	3.13	1.19	0.59	48.06	0.72	0.80	13564.36
CTA	0.92	0.22	2.32	0.89	0.52	56.63	0.85	1.10	-7623.50
CTC	1.27	0.30	3.04	1.19	0.48	37.99	0.93	1.13	-4222.97
CTG	1.61	0.34	2.92	1.61	0.54	33.30	0.90	0.93	7467.87
CTT	1.70	0.49	3.02	1.76	0.53	30.96	0.93	1.01	-4769.33
GAA	1.80	0.75	3.44	1.79	0.43	23.84	0.96	1.08	-6890.45
GAC	1.22	0.19	2.96	1.15	0.56	46.11	0.84	1.00	5083.00
GAG	1.27	0.27	3.00	1.19	0.48	37.74	0.93	1.13	-4329.41
GAT	1.75	0.71	2.63	1.78	0.29	16.81	0.97	0.90	6013.06
GCA	1.69	0.39	3.01	1.71	0.48	28.20	0.90	0.85	12202.75
GCC	2.01	0.08	4.88	1.78	1.33	66.14	0.63	0.73	29416.42
GCG	2.05	0.03	7.30	1.58	1.68	82.04	0.47	0.69	37317.97
GCT	1.64	0.32	2.91	1.68	0.40	24.65	0.91	0.82	12547.97
GGA	1.36	0.47	3.07	1.34	0.41	30.07	0.94	1.01	635.61
GGC	2.01	0.06	4.91	1.75	1.33	66.11	0.63	0.73	29434.18
GGG	1.23	0.07	5.25	1.12	0.71	57.78	0.78	0.87	9058.10
GGT	1.40	0.43	2.66	1.42	0.42	30.17	0.90	0.87	9192.77
GTA	1.03	0.23	2.11	1.08	0.40	38.75	0.91	1.08	-5214.19
GTC	1.22	0.20	2.98	1.14	0.56	46.24	0.84	1.00	5127.23
GTG	1.25	0.27	2.36	1.26	0.41	33.02	0.93	1.08	-454.23
GTT	1.47	0.50	2.61	1.52	0.37	25.50	0.95	1.04	-4405.55
TAA	1.69	0.02	6.60	1.50	1.39	82.07	0.73	1.08	-16662.48
TAC	1.03	0.23	2.15	1.07	0.40	38.72	0.91	1.08	-5256.32
TAG	0.92	0.22	2.29	0.89	0.52	56.65	0.85	1.10	-7508.17
TAT	1.69	0.02	8.82	1.54	1.26	74.46	0.67	1.31	-27963.91
TCA	1.65	0.68	2.54	1.70	0.37	22.31	0.96	1.07	-5620.15
TCC	1.36	0.47	3.08	1.34	0.41	30.07	0.94	1.00	878.35
TCG	1.59	0.07	4.21	1.38	0.98	61.54	0.65	0.81	19473.59
TCT	1.43	0.53	2.87	1.36	0.49	34.19	0.92	1.21	-14347.69
TGA	1.65	0.67	2.49	1.69	0.37	22.41	0.96	1.07	-5669.03
TGC	1.69	0.41	2.98	1.70	0.48	28.24	0.90	0.85	12186.90
TGG	1.60	0.56	2.75	1.60	0.42	26.43	0.94	0.98	3455.04
TGT	1.25	0.33	2.13	1.28	0.34	27.23	0.94	1.36	-17395.63
TTA	1.69	0.02	6.57	1.51	1.39	82.15	0.73	1.07	-16504.00
TTC	1.80	0.75	3.40	1.79	0.43	24.02	0.95	1.07	-6560.65
TTG	1.76	0.32	2.79	1.87	0.58	32.92	0.92	1.07	-8128.09
TTT	2.79	0.04	10.87	2.53	1.97	70.72	0.78	1.21	-37465.94

[a] minimum; [b] maximum; [c] standard deviation; [d] coefficient of variation; [e] Pearson correlation coefficient for the relationship between the observed counts and the expected counts (with all the *P* values < 10^{-26}); [f] slope or intercept of the best-fitted line for the observed counts vs. the expected counts.

There is also a specific pattern for the frequency conservation of trinucleotides across genomes. Among the frequency-conserved trinucleotides, ATC, ATG, CAT, and GAT (all contain AT dinucleotide) are the four most conserved ones across genomes (according to the coefficients of variation). The differences among the means of the frequencies of some conserved trinucleotides are apparent as well (Fig. 2 and Table 2).

3.4. Tetra-, penta-, and hexanucleotide frequencies across genomes

For clarity we present only the data, revealed by the correlation/regression analysis, of the tetranucleotides with the most conserved frequencies across the genomes (Table 3). All these tetranucleotides contain two strong nucleotides and two weak nucleotides, having two G's or C's.

Table 3 Statistical analysis of tetranucleotides with the most conserved frequencies across archaeal, bacterial, and eukaryotic genomes (458 genomes)

Tetranucleotide	Mean (%)	Min[a] (%)	Max[b] (%)	Median (%)	*s* [c]	CV[d] (%)	*r* [e]	Slope[f]	Intercept[f]
ACCA	0.41	0.15	0.69	0.41	0.10	24.31	0.95	1.00	282.88
ACCT	0.32	0.12	0.61	0.32	0.07	21.77	0.96	0.96	526.21
AGGT	0.32	0.13	0.61	0.32	0.07	21.80	0.96	0.96	554.25
ATGG	0.40	0.16	0.69	0.41	0.09	23.03	0.95	0.98	350.23
CCAT	0.40	0.17	0.71	0.41	0.09	23.02	0.95	0.97	505.59
CTTC	0.49	0.19	1.05	0.48	0.12	24.38	0.95	0.98	403.96
GAAG	0.49	0.15	1.06	0.48	0.12	24.37	0.96	0.99	344.95
TGGT	0.41	0.15	0.66	0.41	0.10	24.29	0.95	1.00	299.21

[a] minimum; [b] maximum; [c] standard deviation; [d] coefficient of variation; [e] Pearson correlation coefficient for the relationship between the observed counts and the expected counts (with all the *P* values < 10^{-15}); [f] slope or intercept of the best-fitted line for the observed counts vs. the expected counts.

The obtained results for pentanucleotides show also that some of them are conserved across the genomes in occurrence frequencies. Among these pentanucleotides, the most conserved ones include: ACATC, AGTTC, ATCAC, ATGAC, GAACT, GATGT, GTCAT, GTGAT, GTTCA, and TGAAC. All of them contain two strong nucleotides and three weak nucleotides, having the most conserved dinucleotide AC or GT.

For the hexanucleotides, most of the frequency-conserved ones contain three strong nucleotides and three weak nucleotides; some may contain two strong nucleotides and four weak nucleotides. The hexanucleotides GGAACA, GGTTCA, TGAACC, and TGTTCC are among those with the most conserved frequencies, having also the most conserved dinucleotide AC or GT.

On the whole, the frequencies of hexanucleotides are less conserved across genomes than those of pentanucleotides than those of tetranucleotides than those of trinucleotides than those of dinucleotides according to the coefficient of variation (Table 4). This is consistent with our assumption that the shorter a sequence is, the more chances it would have to be intact during genome evolution. Accordingly, the conservation of occurrence frequencies of higher-order oligonucleotides may become less obvious.

Table 4 Minimums, maximums, and means of the coefficients of variation for frequencies of all di-, tri-, tetra-, penta-, or hexanucleotides across archaeal, bacterial, and eukaryotic genomes (458 genomes)

Short nucleotide	Minimum (%)	Maximum (%)	Mean (%)
Dinucleotide	11.98	66.81	31.85
Trinucleotide	16.74	82.15	42.91
Tetranucleotide	20.58	119.28	51.15
Pentanucleotide	19.57	155.01	59.55
Hexanucleotide	24.13	297.70	67.84

3.5. Analysis of the pattern of strand symmetry

The similarity of the frequency of a mononucleotide, a dinucleotide, or a higher-order oligonucleotide to that of its reverse complement in a genome is very obvious. For mononucleotides the correlation/regression analysis revealed excellent similarities between the count profiles of A and T ($r > 0.999$, $P = 0$, with slope = 1.00, intercept = -837.62) and between the count profiles of C and G ($r > 0.999$, $P = 0$, with slope = 1.00, intercept = 619.32) across all prokaryotic and eukaryotic genomes analyzed. However, the count profiles between A and C, between A and G, between C and T, and between G and T are very different by comparison (r = 0.75, with a slope of 0.91 or 0.61, and a very large intercept). The *t*-test can well distinguish these two kinds of correlation/regression results (for *r*, slope, and intercept, respectively, $P < 0.001$).

The correlation/regression analysis shows also that the count profile across all the genomes of a dinucleotide is different from the count profiles of other dinucleotides except that of its reverse complement (for a dinucleotide and its reverse complement, $r > 0.999$, with a slope of 1.00, and a very small intercept). On the other hand, putting aside the situation of a dinucleotide and its reverse complement, the differences between the count profiles of dinucleotides of the same GC content (e.g., dinucleotides with a GC content of 50% include AC, AG, CA, CT, GA, GT, TC, and TG) are generally smaller compared with those between the count profiles of dinucleotides with different GC content. This is mainly because the expected counts of dinucleotides of the same GC content, as estimated from the frequencies of their component nucleotides, would be the same or very similar in a genome. In spite of this characteristic of dinucleotides of the same GC content (excluding dinucleotides and their respective reverse complements), their correlation/regression results are significantly different from those for dinucleotides and their respective reverse complements according to the *t*-test (for *r*, slope, and intercept, respectively, $P < 0.001$). Apparently, the correlation/regression results for dinucleotides with different GC content would be more different from those for dinucleotides and their respective reverse complements.

Actually the same results were obtained for tri- and tetranucleotides. Most notably, the count profile across all the genomes of a trinucleotide is very similar to that of its reverse complement ($r > 0.999$, with a slope of 1.00, and a very small intercept), while the comparisons between all the other trinucleotide count profiles across the genomes revealed considerable differences (the differences between the count profiles of trinucleotides of the same GC content are also smaller than those between the count profiles of trinucleotides with different GC content, but their correlation/regression results are always significantly different from those for trinucleotides and their respective reverse complements: for *r*, slope, and intercept, respectively, $P < 0.001$). Moreover, the count profile across all the genomes of a tetranucleotide is very similar to that of its reverse complement (several combinations may give somewhat unsatisfied results, e.g., ATTC:GAAT, CCTA:TAGG, CTAA:TTAG, and TCCA:TGGA, with a slope of 0.99, 1.02, 1.01, and 0.99, respectively, and an intercept of 140.55, -97.66, -84.70, and 210.83, respectively), while the comparisons between all the other tetranucleotide count profiles across the genomes revealed considerable differences. Likewise, the *t*-test can well distinguish these two kinds of correlation/regression results (for *r*, slope, and intercept, respectively, $P < 0.001$).

Penta- and hexanucleotides also show similar pattern of strand symmetry. The count profile across all the genomes of a penta- or hexanucleotide is very similar to that of its reverse complement (a few combinations may give somewhat unsatisfied results); the comparisons between all the other penta- or hexanucleotide count profiles across the genomes revealed considerable differences. The *t*-test can well distinguish these two kinds of correlation/regression results (for *r*, slope, and intercept, respectively, $P < 0.001$).

Apparently, the similarity between the count profile across genomes of an oligonucleotide and that of its reverse complement is not only obvious, but also unique. No pairs of count profiles for mono-, di-, and trinucleotides, and few, if any, pairs of count profiles for higher-order oligonucleotides that are not in this category are comparable to them. Thus, strand symmetry is a very explicit phenomenon.

3.6. Characteristics of organelle genomes, plasmids, viruses, phages, and viroids

Most of these genomes or sequences are quite small in size. Nevertheless, results of the analysis of frequency conservation for di- and trinucleotides show that many compositional features found in prokaryotic and eukaryotic genomes are also present, at least for dinucleotides and to some extent, in almost all these genetic systems. In relatively large genomes or long sequences, such as plant mitochondrial genomes and chloroplast genomes, the characteristics are more discernible. It is also worth to note that across genetic systems as small as eukaryotic plasmids, the frequencies of dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG are still more conserved than others.

As for the phenomenon of strand symmetry, it is not conspicuous in these genetic systems. Moreover, in very small genomes or short sequences (e.g., animal mitochondrial genomes and viroids), the phenomenon is almost absent. These results are consistent with those obtained by Mitchell and Bridge (2006), and by Nikolaou and Almirantis (2006).

4. Discussion

4.1. On the causes of oligonucleotide frequency conservation across genomes

The two most important results we obtained from our analysis are: (i) the conservation of frequencies of some di-, tri-, and higher-order oligonucleotides across genomes; and (ii) the phenomenon and pattern of strand symmetry in cellular genomes. The first result has not been reported explicitly by others, at least not in our way and not aiming at finding relics of the primary genetic information (Zhang and Yang, 2005). The phenomenon of strand symmetry, on the other hand, has been a more or less established fact with quite a few studies (see Qi and Cuticchia, 2001; Baisnée et al., 2002; Forsdyke and Bell, 2004; Albrecht-Buehler, 2006). The confirmation in our study of this phenomenon shows that the result on frequency conservation obtained with the same approach is reliable.

The compositional features we reported are universal in archaeal genomes, bacterial genomes, and eukaryotic genomes, no matter what the proportion of non-coding sequences is in a genome. Even *Candidatus Carsonella ruddii* PV and *P. falciparum* (with the lowest GC content in the analyzed prokaryotic genomes and eukaryotic genomes, respectively), and *G. theta* nucleomorph (a vestigial nucleus of eukaryotic endosymbiont) are compatible quite well with the general regime. In fact, we got virtually the same results with new whole-genome sequences added in the analysis (our unpublished data).

Early study indicates that there are significant correlations between genomic libraries in terms of tetranucleotide frequency distribution, suggesting an overall correlation of frequency profiles of short nucleotides among genomes (Rogerson, 1991). Our finding shows that the frequency conservation involves only some di-, tri-, and higher-order oligonucleotides, especially the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG. It may be true that many individual mononucleotides have not changed in the course of billions of years of evolution. However, the genomic frequencies of mononucleotides are not well conserved, in concordance with the fact that only half of the 16 dinucleotides are well conserved in genomic frequencies. Causes for this phenomenon may include: (i) patterns of distributions of short nucleotides throughout a genome and across genomes; and (ii) probabilities of occurrences of short nucleotides set by strand symmetry.

It has been shown that genome inhomogeneity is determined mainly by AA, TT, GG, CC, AT, TA, GC and CG dinucleotides, which are closely associated with polyW and polyS tracts (W and S stand for weak nucleotides and strong nucleotides, respectively) (Kozhukhin and Pevzner, 1991). This implies that the distribution of any one of the other eight dinucleotides (SW and WS dinucleotides, i.e., AC, AG, CA, CT, GA, GT, TC, and TG) in a genome is rather homogeneous. Also, the distributions of oligonucleotides containing similar and especially the same numbers of the strong and weak nucleotides, but no CG or TA dinucleotide, are the most uniform in six representative genomes (yet the authors considered their distributions not informative) (Häring and Kypr, 1999). The results of our analysis are consistent with these distribution patterns, except that oligonucleotides with well conserved frequencies may contain TA dinucleotide. Moreover, when the frequency profiles of different chromosomes of a species are not very similar (e.g., *D. melanogaster* and *H. sapiens*), the dinucleotides with well conserved frequencies across genomes have also more conserved frequencies across chromosomes than other dinucleotides (our unpublished results). Results of further analysis with a sliding window of 1 kb indicate that the distributions of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides are indeed much more homogeneous across genomes than other dinucleotides (Zhang and Lan, unpublished results). Therefore, one reason for the frequency conservation of some di-, tri-, and higher-order oligonucleotides across genomes would be that their distributions are quite uniform throughout a genome and across genomes.

In addition, if the probability of occurrences of a short nucleotide is fixed to a certain range by the frequencies of the component nucleotides (which themselves follow the rule of strand symmetry), the variation of its actual frequency will also be limited. For example, in our analyzed prokaryotic and eukaryotic genomes, the GC content varies from 16.6% to 74.9% (with the percentage of A + T varying from 25.1% to 83.4%). Under the regime of strand symmetry, the expected frequencies of AA, AT, TA, and TT dinucleotides may vary from 1.6% (with the frequencies of A and T being both approximately 12.6%) to 17.4% (with the frequencies of A and T being both approximately 41.7%), and those of CC, CG, GC, and GG dinucleotides from 0.7% (with the frequencies of C and G being both approximately 8.3%) to 14.0% (with the frequencies of C and G being both approximately 37.5%). However, the expected frequencies of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides will range only from 3.5% (GC content being 16.6%) to 6.3% (GC content being 50.0%). Therefore, strand symmetry would contribute to frequency conservation; it is not unusual that AC, AG, CA, CT, GA, GT, TC, and TG are the dinucleotides with well-conserved frequencies across modern genomes. Also as a result of strand symmetry, given that the frequencies of a short nucleotide are conserved across genomes, so would be those of its reverse complement, doubling the number of different frequency-conserved short nucleotides. Thus, the contribution of strand symmetry to the frequency conservation would be twofold: limiting the range of variation and doubling the content of conservation.

The two causes for the frequency conservation of some di-, tri-, and higher-order oligonucleotides across genomes would be somewhat correlated: the rule of strand symmetry itself may influence the patterns of distributions of short nucleotides throughout a genome and across genomes. Nevertheless, strand symmetry would not be the only cause for frequency conservation. This seems apparent because strand symmetry alone could not fully explain the patterns of frequency conservation of dinucleotides, trinucleotides (see also Albrecht-Buehler, 2007a), and higher-order oligonucleotides.

4.2. On the origin of frequency conservation and strand symmetry: Constitution of the primordial genome

The origin of the patterns of frequency conservation and strand symmetry would in no way be by random processes. We discuss only the situations of random sequences with characteristics related to frequency conservation or strand symmetry. First, in random and sufficiently long (e.g., with lengths comparable to cellular genome sizes) sequences with fixed probability of occurrences for each component nucleotide (e.g., A = 25%, C = 20%, G = 30%, and T = 25% in all these sequences), the frequencies of every mono-, di-, or higher-order oligonucleotide would tend to be equal to its probability. Thus, frequencies of every oligonucleotide would be conserved or even constant across such sequences. When all component nucleotides are not with fixed probabilities of occurrences, there would be various patterns of frequency conservation and variation across the long random sequences (e.g., if only A with fixed probability, then for oligonucleotides of the same order, only those composed entirely of A would be with conserved or constant frequencies). Second, in a long random sequence generated according to the rule of first-order symmetry (strand symmetry for mononucleotides), the frequencies of oligonucleotides of the same GC content (see section 3.5 for more details) and of the same order would be very similar. Indeed, the frequencies of oligonucleotides equal those of their reverse complements and also those of their forward complements in such a sequence (Qi and Cuticchia, 2001). In general, the observed frequencies of short nucleotides in a genome and those expected from the frequencies of their component nucleotides would be moderately, if not highly, correlated. If there were only first-order symmetry in genomic sequences, their characteristics of oligonucleotide frequencies would be more or less similar to those of the random sequence mentioned above.

Our study indicates that there are specific patterns for the frequency conservation of oligonucleotides across genomes. Similar patterns could not be found in random sequences without first-order symmetry (in the case of random sequences with first-order symmetry, it is the second situation discussed above). And although there are traces of similarity for the frequencies of oligonucleotides of the same GC content and of the same order (especially higher-order oligonucleotides, putting aside the situation of an oligonucleotide and its reverse complement, see also section 3.5) in a genome, only the similarity between the frequency of an oligonucleotide and that of its reverse complement is very obvious. In addition, the phenomenon of strand symmetry persists for oligonucleotides up to 10 nt long at least (Qi and Cuticchia, 2001). Therefore, the phenomenon of frequency conservation of particular di-, tri-, and higher-order oligonucleotides across genomes and the pattern of first-order and high-order strand symmetries would be characteristics unique to genomic sequences. It is highly improbable for long sequences with first-order symmetry only to become natural genomic sequences by random shuffling (see also Forsdyke, 1995b; Forsdyke and Bell, 2004).

To find out the origin of the universal features of frequency conservation and strand symmetry, there would be two alternative approaches. One approach is to consider the universal features as evolutionary convergences (the current hypotheses on the origin of strand symmetry seem to focus on this approach, see below); the other is to consider the universal features as relics (original characteristics) of the primordial genome. The first one has to reveal the universal selective advantages or mutation pressures leading to the convergences, which are not always apparent (see below). Therefore, we tried the alternative: considering the universal features as relics of the primordial genome.

No matter whether the compositional features in modern genomes are due to structural constraints or other factors on nucleic acid sequences, the constraints or factors might exist from the very beginning of genome evolution. Under this consideration, the compositional features would be relics rather than convergences. In any case, considering the universal features as “relics” would be a more economic explanation than the “evolutionary convergences” point of view.

Why are there high-order symmetries besides first-order symmetry? Several explanations for the origin of strand symmetry have been proposed, such as no strand biases for mutation and selection (for first-order symmetry only, see Sueoka, 1995; Lobry, 1995; Lobry and Lobry, 1999), strand inversion (Fickett et al., 1992; see also Albrecht-Buehler, 2006 for inversion and inverted transposition), selection of stem-loop structures (Forsdyke, 1995a; 1995b; Forsdyke and Bell, 2004), and combined effects of a wide spectrum of mechanisms operating at multiple orders and length scales (Baisnée et al., 2002). However, it seems that the fundamental cause of strand symmetry is still unclear (see also Baisnée et al., 2002; Chen and Zhao, 2005). The hypothesis of stem-loop mechanism argues that stem-loop structures would be advantageous for recombination, so that mutations favoring the general potential of the formation of stem-loop structures in single-stranded DNA (hence strand symmetry) would confer a selective advantage (Bell and Forsdyke, 1999; Forsdyke and Mortimer, 2000). However, there are cases indicating that local recombination rates would be negatively correlated with levels of higher-order strand symmetry (Chen and Zhao, 2005). Even the latest quantitative transposition/inversion model (Albrecht-Buehler, 2006) could not accommodate the patterns of conservation of oligonucleotide frequency profiles if the initial profiles were arbitrary ones (yet it would be a good model for the maintenance of strand symmetry, see also Albrecht-Buehler, 2007b). Moreover, the relation of strand symmetry to the phenomenon of frequency conservation has not been emphasized in the current hypotheses.

Alternatively, considering the universal compositional features as relics of the primordial genome, the shared features must have arisen very early in evolution. We have now evidences indicating that uniform distribution of some di-, tri-, and higher-order oligonucleotides would be one of the causes of the phenomenon of frequency conservation. Therefore, the ancestor of modern genomes — the primordial genome — would also have this property, i.e., compositional homogeneity, at least for some di-, tri-, and higher-order oligonucleotides, throughout the genome. What kind of structure, apart from random sequences, has this property? A very possible one is repeated sequences. It has been proposed that repeats of a kind of macromolecules capable of self-replicating made up the most primitive genes and genomes (Zhang, 1998a; 1998b). When these DNA (RNA) macromolecules formed repeated sequences by connecting one after another, there would be no force preventing them from producing approximately equal amounts of forward repeats and their reverse repeats, leading naturally to strand symmetry. Sequences formed in this way would consequently have high potentials for perfect stem-loop structures, especially in the early stages of evolution (see also Forsdyke, 1996).

Based on the above considerations, we have a hypothesis for the origin of frequency conservation and first-order and high-order strand symmetries. The primordial genome would be composed of repeated sequences: approximately equal amounts of rather uniformly distributed forward repeats and their reverse repeats. Also, the compositional features of the primitive repeating units would be reflected to some extent from the specific patterns of oligonucleotide frequency conservation in modern genomes. The universal compositional features found in modern genomes would therefore have an origin in the primordial genome. This proposition is simple and efficient in explaining various characteristics of oligonucleotide frequencies in modern genomes. Given that strand symmetry has existed from the very beginning, the T vs. A and G vs. C skews due to strand-dependent mutations in the leading and the lagging halves of bacterial genomes may in principle be canceled out over the whole genome, which is exactly the case (Shioiri and Takahata, 2001). Also, if strand symmetry has existed from the beginning, the frequency variations across genomes of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides and some higher-order oligonucleotides would be limited, no matter what the GC content of the primordial genome was. In other words, if strand symmetry is a relic of the primordial genome, so must be the pattern of frequency conservation linked to it.

The phenomenon of strand symmetry is more obvious than that of frequency conservation (according to the correlation/regression analysis). Also, strand symmetry would be one of the main causes for frequency conservation. Therefore, strand symmetry would be a primary feature. Whether frequency conservation would also be a primary feature or just a secondary consequence to strand symmetry is a matter of further study. On the other hand, as frequency conservation is of specific patterns that could not be explained by strand symmetry alone, it is possible that the patterns would also be primary features.

The maintenance of strand symmetry during evolution would rely on multiple mechanisms such as duplication and rearrangement (inversion), or more directly, inverse duplication (Nussinov, 1982; Sanchez and Jose, 2002) or inverted transposition/inversion (Albrecht-Buehler, 2006). In turn, strand symmetry would be the system (mechanism) to maintain frequency conservation. In organelle genomes and some other genetic systems, the phenomena of frequency conservation and strand symmetry are less obvious or even absent. The causes would be that these genetic systems are very specified (e.g., specific loss of genome material during evolution), too small, or devoid of the mechanisms to maintain strand symmetry or frequency conservation. For example, it appears that rearrangements have occurred relatively infrequently in animal mitochondrial genomes (Gray, 1989).

In addition to the facts and arguments presented previously (Zhang, 1998a; 1998b), we now have further evidences relating to the compositional features of modern genomes for the hypothesis that the most primitive nucleic acid genome would be composed of repeated sequences. We propose further that the primary genetic information would be rather uniformly distributed throughout the primordial genome in the form of direct and inverted repeated sequences. It would now become a kind of “genomic background” upon which genome evolution takes place. The primary genetic information, which may also be termed “the primary genetic code”, would have its relics in modern genomes: strand symmetry and frequency conservation. These relics may just be the “highly conserved pattern underlying all other genetic information in cellular DNA” suggested by Rogerson (1991). The primary genetic information revealed from modern structures would certainly help us to reconstruct the primordial genome as well as to understand the patterns and processes of genome evolution, thus would shed light on the origin and evolution of genomes, and even on the origin of life. Apparently, the information may also be used in the study of molecular systematics based on whole-genome sequences.

Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (No. 30270752) and a grant from the Guangdong Natural Science Foundation (No. 031616). Also, it was supported by the National Basic Research Program (No. 2005CB724600).

References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., Ikemura, T., 2003. Informatics for unveiling hidden genome signatures. Genome Res. 13, 693–702.

Albrecht-Buehler, G., 2006. Asymptotically increasing compliance of genomes with Chargaff’s second parity rules through inversions and inverted transpositions. Proc. Natl. Acad. Sci. USA 103, 17828–17833.

Albrecht-Buehler, G., 2007a. The three classes of triplet profiles of natural genomes. Genomics 89, 596–601.

Albrecht-Buehler, G., 2007b. Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences. Genomics 90, 297–305.

Baisnée, P.F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? Bioinformatics 18, 1021–1033.

Bell, S.J., Forsdyke, D.R., 1999. Accounting units in DNA. J. Theor. Biol. 197, 51–61.

Bohlin, J., Skjerve, E., Ussery, D.W., 2008. Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. BMC Genomics 9, 104.

Burge, C., Campbell, A.M., Karlin, S., 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA 89, 1358–1362.

Chargaff, E., 1951. Structure and function of nucleic acids as cell constituents. Fed. Proc. 10, 654–659.

Chen, L., Zhao, H., 2005. Negative correlation between compositional symmetries and local recombination rates. Bioinformatics 21, 3951–3958.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature:

characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.

Fickett, J.W., Torney, D.C., Wolf, D.R., 1992. Base compositional structure of genomes. Genomics 13, 1056–1064.

Forsdyke, D.R., 1995a. A stem-loop “kissing” model for the initiation of recombination and the origin of introns. Mol. Biol. Evol. 12, 949–958.

Forsdyke, D.R., 1995b. Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J. Mol. Evol. 41, 573–581.

Forsdyke, D.R., 1996. Different biological species “broadcast” their DNAs at different (G + C)% “wavelengths”. J. Theor. Biol. 178, 405–417.

Forsdyke, D.R., Mortimer, J.R., 2000. Chargaff's legacy. Gene 261, 127–137.

Forsdyke, D.R., Bell, S.J., 2004. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl. Bioinformatics 3, 3–8.

Gray, M.W., 1989. Origin and evolution of mitochondrial DNA. Annu. Rev. Cell Biol. 5, 25–50.

Häring, D., Kypr, J., 1999. Variations of the mononucleotide and short oligonucleotide distributions in the genomes of various organisms. J. Theor. Biol. 201, 141–156.

Karkas, J.D., Rudner, R., Chargaff, E., 1968. Seapration of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl. Acad. Sci. USA 60, 915–920.

Karlin, S., Ladunga, I., Blaisdell, B.E., 1994. Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA 91, 12837–12841.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.

Karlin, S., Mrazek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.

Kozhukhin, C.G., Pevzner, P.A., 1991. Genome inhomogeneity is determined mainly by WW and SS dinucleotides. Comput. Appl. Biosci. 7, 39–49.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Inc., Sunderland, MA, pp. 177–196.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40, 326–330.

Lobry, J.R., Lobry, C., 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol. 16, 719–723.

Mitchell, D., Bridge, R., 2006. A test of Chargaff’s second rule. Biochem. Biophys. Res.

Commun. 340, 90–94., doi:10.1016/j.bbrc.2005.11.160.

Nikolaou, C., Almirantis, Y., 2006. Deviations from Chargaff's second parity rule in organellar DNA: insights into the evolution of organellar genomes. Gene 381, 34–41., doi:10.1016/j.gene.2006.06.010.

Nussinov, R., 1980. Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res. 8, 4545–4562.

Nussinov, R., 1981. Nearest neighbor nucleotide patterns: structural and biological implications. J. Biol. Chem. 256, 8458–8462.

Nussinov, R., 1982. Some indications for inverse DNA duplication. J. Theor. Biol. 95, 783–791.

Nussinov, R., 1984. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res. 12, 1749–1763.

Okamura, K., Wei, J., Scherer, S.W., 2007. Evolutionary implications of inversions that have caused intra-strand parity in DNA. BMC Genomics 8, 160.

Prabhu, V.V., 1993. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 21, 2797–2800.

Qi, D., Cuticchia, A.J., 2001. Compositional symmetries in complete genomes. Bioinformatics 17, 557–559.

Rogerson, A.C., 1989. The sequence asymmetry of the *Escherichia co1i* chromosome appears to be independent of strand or function and may be evolutionarily conserved. Nucleic Acids Res. 17, 5547–5563.

Rogerson, A.C., 1991. There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. J. Mol. Evol. 32, 24–30.

Rudner, R., Karkas, J.D., Chargaff, E., 1968. Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis. Proc. Natl. Acad. Sci. USA 60, 921–922.

Sanchez, J., Jose, M.V., 2002. Analysis of bilateral inverse symmetry in whole bacterial chromosomes. Biochem. Biophys. Res. Commun. 299, 126–134.

Shioiri, C., Takahata, N., 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. J. Mol. Evol. 53, 364–376.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48, 582–592.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325.

van Passel, M.W., Kuramae, E.E., Luyf, A.C., Bart, A., Boekhout, T., 2006. The reach of the genome signature in prokaryotes. BMC Evol. Biol. 6, 84.

Wang, Y., Hill, K., Singh, S., Kari, L., 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene 346, 173–185.

Watson, J.D., Crick, F.H.C., 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171, 737–738.

Zhang, S.-H., 1998a. The origin and evolution of repeated sequences and introns. Speculat. Sci. Technol. 21, 7–16.

Zhang, S.-H., 1998b. Origin and evolution of exon/intron junctions. Speculat. Sci. Technol. 21, 17–27.

Zhang, S.-H., Yang, J.-H., 2005. Conservation versus variation of dinucleotide frequencies across genomes: evolutionary implications. Genome Biol. 6, P12., doi:10.1186/gb-2005-6-11-p12 (deposited research, http://genomebiology.com/2005/6/11/P12).