Deposited research article

Conservation versus variation of dinucleotide frequencies across genomes: Evolutionary implications

Shang-Hong Zhang*, Jian-Hua Yang

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

Correspondence: Shang-Hong Zhang. Email: lsszsh@zsu.edu.cn

Posted: 11 October 2005

Genome ***Biology*** 2005, **6:**P12

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/11/P12

Received: 6 October 2005

This is the first version of this article to be made available publicly.

Conservation versus variation of dinucleotide frequencies across genomes: Evolutionary implications

Shang-Hong Zhang*, Jian-Hua Yang

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

Corresponding author: Shang-Hong Zhang

Address: Biotechnology Research Center, Sun Yat-Sen University, Guangzhou 510275, China

Tel: 86-20-84110316; 86-20-84035425

Email: lsszsh@zsu.edu.cn

Running head: Dinucleotide frequencies across genomes

Abstract

Background

In order to find traits or evolutionary relics of the primordial genome (the most primitive nucleic acid genome for earth's life) remained in modern genomes, we have studied the characteristics of dinucleotide frequencies across genomes. As the longer a sequence is, the more probable it would be modified during genome evolution. For that reason, short nucleotide sequences, especially dinucleotides, would have considerable chances to be intact during billions of years of evolution. Consequently, conservation of the genomic profiles of the frequencies of dinucleotides across modern genomes may exist and would be an evolutionary relic of the primordial genome.

Results

Based on this assumption, we analyzed the frequency profiles of dinucleotides of the whole-genome sequences from 130 prokaryotic species (including archaea and bacteria). The statistical results show that the frequencies of the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG are well conserved across genomes, while the frequencies of other dinucleotides vary considerably among species. This conservation/variation seems to be linked to the distributions of dinucleotides throughout a genome and across genomes, and also to have relation to strand symmetry.

Conclusions

We argue and conclude that the phenomenon of frequency conservation would be evolutionary relics of the primordial genome, which may provide insights into the study of the origin and evolution of genomes.

Key words: Dinucleotide frequencies — Compositional analysis — Whole-genome sequences — Strand symmetry — Primordial genome — Evolutionary relics — Origin and evolution of genomes

Background

In the course of billions of years of evolution, organic genomes have undergone enormous changes. Nevertheless, some traits or evolutionary relics of the primordial genome (defined as the most primitive nucleic acid genome for earth's life in this paper) may remain in modern genomes. Finding these traits or relics is of great significance for the study of the origin and evolution of genomes [1]. Indeed, the only way to reconstitute ancient genomes in the absence of fossil DNA may be the deduction from the comparative analysis of the structures of present-day genomes [2].

What traits at the genomic level may be regarded as evolutionary relics of the primordial genome? One consideration is that for a sequence of DNA (or RNA) in a genome, the longer it is, the more probable it would be modified (such as nucleotide substitutions, insertions or deletions, but not including duplication) during genome evolution. For that reason, the shortest possible sequences, the dinucleotides, would have in general the most chances to be intact as pieces of sequence. If a considerable proportion of a particular dinucleotide was intact in evolving genomes, its genomic occurrence frequencies would not change significantly during genome evolution. Based on this assumption, comparative analysis of the characteristics of dinucleotides in the genomes of various organisms may provide insights into the features of the primordial genome as well as the genetic information it contained.

From this point forward, our philosophy suggests that the conservation of the genomic profiles of the frequencies of dinucleotides across various genomes, if it exists, would be an evolutionary relic of the primordial genome. In other words, if the frequencies of a

dinucleotide in modern genomes are conserved, it would imply that the genomic frequencies of that particular dinucleotide have not changed significantly since the primordial genome formed. For mononucleotides, it has been known that their frequencies vary among species, especially in prokaryotes [3]. However, this does not preclude the possibility of the conservation of the frequencies of some, if not all, dinucleotides across genomes. Many researches have been done in the field of dinucleotide frequencies even when sequence data were limited (e.g., [4, 5, 6]), 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 this field is the characteristics of dinucleotide relative abundances, which access contrasts between the observed dinucleotide frequencies and those expected from the component nucleotide frequencies [7]. The profiles of relative abundances of dinucleotides in genomic sequences are rather species-specific or taxon-specific [8, 9]. The set of all dinucleotide relative abundance values is even regarded as a genomic signature [7]. This specificity seems in contradiction with our assumption on the conservation of the frequency profiles. 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 [10]. Moreover, instead of considering the frequencies of all dinucleotides 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 dinucleotide occurrence frequencies exists across genomes, or to determine to what extent the frequencies of a dinucleotide vary among species.

With the development of genomics, more and more whole-genome sequences are now available, providing opportunities for the analysis of evolutionary relics at the genomic level. In this paper we analyzed the frequency profiles of dinucleotides of the whole-genome sequences from 130 species (including archaea and bacteria). The results show that the conservation of frequencies of some dinucleotides across genomes does exist. We argue and conclude that the frequency conservation would be evolutionary relics of the primordial genome.

Results and Discussion

The distribution pattern of the frequencies of 16 dinucleotides of 130 species of archaea and bacteria is shown in Figure 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. While the distributions of the frequencies of AA, AT, CC, CG, GC, GG, TA, and TT dinucleotides are dispersed throughout their respective ranges, most of the genomic frequencies of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides are clustered around their own means (see also Additional File 1 for the details of the results). This characterization is also evident from the statistics such as the standard deviation, the coefficient of variation, the minimum and the maximum of the dinucleotide frequencies (Table 1). As for the dinucleotide counts, the correlation coefficients for the observed vs. expected counts of the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG are very close to 1 ($P < 10^{-50}$), with also the slopes close to 1 and

the intercepts relatively small. 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. For the other eight dinucleotides (AA, AT, CC, CG, GC, GG, TA, and TT), the correlation coefficients are between 0.32 ($P < 10^{-3}$) and 0.88 ($P < 10^{-36}$), but the slopes are not close to 1 and the intercepts are relatively large. Furthermore, the χ^2 test revealed no significant difference in terms of average counts/kb across the archaeal and bacterial genomes for AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides only ($P > 0.05$) (Table 1, see also Additional File 1 for details). The χ^2 test revealed also among the eight frequency-conserved dinucleotides, AC and GT are with the largest *P* values, and AG and CT are with the smallest *P* values. 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 (a phenomenon that reflects the similarities of the frequencies of nucleotides and oligonucleotides to those of their respective reverse complements within single strands of genomic sequences, see [11, 12, 13, 14]).

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, hence somewhat trivial. 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 conservation across genomes of the dinucleotides concerned.

Both the correlation/regression analysis and the χ^2 test indicate that the frequencies of the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG are well conserved across genomes, while the frequencies of the dinucleotides AA, AT, CC, CG, GC, GG, TA, and TT vary considerably among species. The χ^2 test seems give more clear-cut results, hence would be a good and simple choice for testing the conservation of dinucleotide frequencies (for the appropriateness of the χ^2 test, see also [14]). Though our data concern only prokaryotic genomes, actually very similar results were obtained when the sequences of the prokaryotic genomes and almost all the currently available complete eukaryotic genomes were taken together in the χ^2 test (our unpublished results). Therefore, the conservation of the frequencies of some dinucleotides across genomes does exist. These results have not been reported so far, at least not in our way and not aiming at finding evolutionary relics of the primordial genome. In fact, the conservation of the frequencies of the dinucleotides AC, AG, CA, CT, GA, GT, TC, and TG is more observable than that of any mononucleotide, trinucleotide, or higher-order oligonucleotide (our unpublished results). 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 terms of genomic frequencies.

The conservation of the frequencies of some dinucleotides we reported is universal in

modern archaeal genomes and bacterial genomes. Also, it is found, with evidences, in eukaryotic genomes, even if they have a large proportion of non-coding sequences. To explain the existence of these universal features, one reasonable approach would be to consider them as the evolutionary relics of the primordial genome. No matter whether these compositional features are due to structural constraints or other factors on nucleic acid sequences, the constraints or factors, probably chemical or physical in nature, would exist from the very beginning of genome evolution. Thus, the compositional features would be evolutionary relics rather than convergences.

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 [15]. Our finding shows that the frequency conservation involves especially some dinucleotides. Causes for this phenomenon may include: (i) patterns of distributions of dinucleotides throughout a genome and across genomes; and (ii) probabilities of occurrences of dinucleotides 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 (consisting of two strong nucleotides or two weak nucleotides), which are closely associated with polyW and polyS tracts (W and S stand for weak nucleotides and strong nucleotides, respectively) [16]. 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) [17]. The results of our analysis are consistent with these distribution patterns. Therefore, one reason for the frequency conservation across genomes of some but not all dinucleotides would be that only the distributions of the frequency-conserved dinucleotides are quite uniform throughout a genome and across genomes.

In addition, if the probability of occurrences of a dinucleotide 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 genomes, the AT content varies from 27.9% to 77.5%; the GC content varies from 22.5% to 72.1%. Under the regime of strand symmetry, the expected frequencies of AA, AT, TA, and TT dinucleotides may vary from 1.9% (with the frequencies of A and T being both approximately 14.0%) to 15.0% (with the frequencies of A and T being both approximately 38.8%), and those of CC, CG, GC, and GG dinucleotides from 1.3% (with the frequencies of C and G being both approximately 11.3%) to 13.0% (with the frequencies of C and G being both approximately 36.1%). However, the expected frequencies of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides will range only from 4.4% (AT content being 77.5%, GC content being 22.5%) to 6.3% (both AT content and 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.

Mononucleotide frequencies would tend to diverge during genome evolution. Therefore, the variation of mononucleotide frequencies among species would be a derived trait. On the other hand, the phenomenon of strand symmetry is ubiquitous and would probably exist from the very beginning (a subject to be discussed in a separate paper). In other words, if strand symmetry is an evolutionary relic of the primordial genome, so must be frequency conservation linked to it. Furthermore, the mechanisms for maintaining strand symmetry, such as inverse duplication [18, 19], would also help to maintain frequency conservation.

Conclusion

The phenomenon of frequency conservation is universal and consistent in modern genomes. These shared compositional features are likely to have arisen very early in evolution. Therefore, we conclude that the phenomenon of frequency conservation would be evolutionary relics of the primordial genome, implying that the primordial genome would have similar frequencies of AC, AG, CA, CT, GA, GT, TC, and TG dinucleotides as modern genomes — probably very close to the mean values calculated from modern genomes. On the other hand, the genomic frequencies of AA, AT, CC, CG, GC, GG, TA, and TT dinucleotides would vary during genome evolution. This kind of information revealed from modern structures would certainly help us to reconstruct the primordial genome as well as to understand the pattern of genome evolution, thus would shed light on the origin and evolution of genomes, and even on the origin of life.

Materials and Methods

Whole-Genome Sequences

We downloaded the whole-genome sequence of every species of archaea and bacteria that was available as of May 2004 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 from each of them (our analysis shows that the choice of samples does not influence the validity of the results, data not shown). In total, 18 complete genomes of archaea and 112 complete genomes of bacteria were analyzed in the study (A denotes an archaeon): *Aeropyrum pernix K1* (A), *Agrobacterium tumefaciens str. C58*, *Aquifex aeolicus VF5*, *Archaeoglobus fulgidus DSM 4304* (A), *Bacillus anthracis Ames*, *B. cereus ATCC 14579*, *B. halodurans C-125*, *B. subtilis subsp. subtilis 168*, *Bacteroides thetaiotaomicron VPI-5482*, *Bdellovibrio bacteriovorus HD100*, *Bifidobacterium longum NCC2705*, *Bordetella bronchiseptica RB50*, *B. parapertussis 12822*, *B. pertussis Tohama I*, *Borrelia burgdorferi B31*, *Bradyrhizobium japonicum USDA 110*, *Brucella melitensis 16M*, *B. suis 1330*, *Buchnera aphidicola str. Bp*, *Campylobacter jejuni subsp. jejuni NCTC 11168*, *Candidatus Blochmannia floridanus*, *Caulobacter crescentus CB15*, *Chlamydia muridarum*, *C. trachomatis D/UW-3/CX*, *Chlamydophila caviae GPIC*, *C. pneumoniae J138*, *Chlorobium tepidum TLS*, *Chromobacterium violaceum ATCC 12472*, *Clostridium acetobutylicum ATCC 824*, *C. perfringens 13*, *C. tetani E88*, *Corynebacterium diphtheriae NCTC 13129*, *C. efficiens YS-314*, *C. glutamicum ATCC 13032*, *Coxiella burnetii RSA 493*, *Deinococcus radiodurans R1*, *Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough*,

Enterococcus faecalis V583, *Escherichia coli K12*, *Fusobacterium nucleatum subsp. nucleatum ATCC 25586*, *Geobacter sulfurreducens PCA*, *Gloeobacter violaceus PCC 7421*, *Haemophilus ducreyi 35000HP*, *H. influenzae Rd*, *Halobacterium sp. NRC-1* (A), *Helicobacter hepaticus ATCC 51449*, *H. pylori 26695*, *Lactobacillus johnsonii NCC 533*, *L. plantarum WCFS1*, *Lactococcus lactis subsp. lactis Il1403*, *Leptospira interrogans serovar lai str. 56601*, *Listeria innocua Clip11262*, *L. monocytogenes EGD-e*, *Mesorhizobium loti MAFF303099*, *Methanobacterium thermoautotrophicum str. deltaH* (A), *Methanococcus jannaschii DSM 2661* (A), *M. maripaludis S2* (A), *Methanopyrus kandleri AV19* (A), *Methanosarcina acetivorans C2A* (A), *M. mazei Goe1* (A), *Mycobacterium avium subsp. paratuberculosis k10*, *M. bovis AF2122/97*, *M. leprae TN*, *M. tuberculosis H37Rv*, *Mycoplasma gallisepticum R*, *M. genitalium G-37*, *M. mycoides subsp. mycoides SC str. PG1*, *M. penetrans HF-2*, *M. pneumoniae M129*, *M. pulmonis UAB CTIP*, *Nanoarchaeum equitans Kin4-M* (A), *Neisseria meningitidis Z2491*, *Nitrosomonas europaea ATCC 19718*, *Nostoc sp. PCC 7120*, *Oceanobacillus iheyensis HTE831*, *Onion yellows phytoplasma OY-M*, *Parachlamydia sp. UWE25*, *Pasteurella multocida Pm70*, *Photorhabdus luminescens subsp. laumondii TTO1*, *Pirellula sp. 1*, *Porphyromonas gingivalis W83*, *Prochlorococcus marinus subsp. marinus str. CCMP1375*, *Pseudomonas aeruginosa PAO1*, *P. putida KT2440*, *P. syringae pv. tomato DC3000*, *Pyrobaculum aerophilum IM2* (A), *Pyrococcus abyssi GE5* (A), *P. furiosus DSM 3638* (A), *P. horikoshii OT3* (A), *Ralstonia solanacearum GMI1000*, *Rhodopseudomonas palustris CGA009*, *Rickettsia conorii str. Malish 7*, *R. prowazekii str. Madrid E*, *Salmonella enterica subsp. enterica serovar Typhi str. CT18*, *S. typhimurium*

LT2, *Shewanella oneidensis MR-1*, *Shigella flexneri 2a str. 301*, *Sinorhizobium meliloti 1021*, *Staphylococcus aureus subsp. aureus Mu50*, *S. epidermidis ATCC 12228*, *Streptococcus agalactiae 2603V/R*, *S. mutans UA159*, *S. pneumoniae R6*, *S. pyogenes MGAS315*, *Streptomyces avermitilis MA-4680*, *S. coelicolor A3(2)*, *Sulfolobus solfataricus P2* (A), *S. tokodaii str. 7* (A), *Synechococcus sp. WH 8102*, *Synechocystis sp. PCC 6803*, *Thermoanaerobacter tengcongensis str. MB4T*, *Thermoplasma acidophilum DSM 1728* (A), *T. volcanium GSS1* (A), *Thermosynechococcus elongatus BP-1*, *Thermotoga maritima MSB8*, *Thermus thermophilus HB27*, *Treponema denticola ATCC 35405*, *T. pallidum subsp. pallidum str. Nichols*, *Tropheryma whipplei Twist*, *Ureaplasma urealyticum ATCC 700970*, *Vibrio cholerae O1 biovar eltor str. N16961*, *V. parahaemolyticus RIMD 2210633*, *V. vulnificus CMCP6*, *Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis*, *Wolbachia endosymbiont of Drosophila melanogaster*, *Wolinella succinogenes DSM 1740*, *Xanthomonas axonopodis pv. citri str. 306*, *X. campestris pv. campestris str. ATCC 33913*, *Xylella fastidiosa Temecula1*, *Yersinia pestis CO92*. For accession numbers of these sequences, see Additional File 1.

Calculations of Genomic Occurrence Frequencies of Dinucleotides

We counted the number of occurrences of every dinucleotide in each genome. The count was performed in all the possible reading frames (equivalent to moving the sliding window of 2 nt down the sequence one base at a time). Each chromosome was analyzed separately, without concatenation. Counts were compiled for each species. 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, only one strand of each genome (the downloaded sequence) was analyzed. Although the choice of strands seems arbitrary, strand symmetry [11, 12, 13, 14] guarantees the validity of the results. In fact, there is little difference in terms of dinucleotide occurrence frequencies in analyzing one strand or another or both strands of a genome (data not shown).

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

Statistical Analysis

To test whether the frequencies of dinucleotides are conserved across genomes, we employed the correlation/regression analysis followed by a χ^2 test. For each dinucleotide, we analyzed the correlation between the observed counts and the expected counts in the genomes studied. The expected count of a dinucleotide in a genome was obtained from the total of all dinucleotide counts of that genome multiplied by the mean frequency of that particular dinucleotide (the average of all genomes studied). We calculated the correlation coefficient (*r*), the slope and intercept of the best-fitted line for the observed counts vs. the expected counts. As for the χ^2 test, we employed it at a less stringent level because the analysis is to reveal relics that would be "buried" and "dissolved" in modern structures. Instead of using the total counts in a genome, we used the average counts/kb. Therefore, in the χ^2 test the observed value of a particular dinucleotide of a species was its occurrence frequency (percentage) multiplied by 1,000 and rounded to an integer. The expected value was the average of the corresponding observed values of all the species concerned.

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).

References

1. Zhang S-H: **On the origin and evolution of organic genomes.** *Acta Scientiarum Naturalium Universitatis Sunyatseni* 1996, **35:**96–101.

2. Birnbaum D, Coulier F, Pebusque MJ, Pontarotti P: **"Paleogenomics": Looking in the past to the future.** *J Exp Zool* 2000, **288:**21–22.

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

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

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

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

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

8. Karlin S, Ladunga I, Blaisdell BE: **Heterogeneity of genomes: Measures and values.** *Proc Natl Acad Sci USA* 1994, **91:**12837–12841.

9. Karlin S, Mrazek J, Campbell AM: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179:**3899–3913.

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

11. Fickett JW, Torney DC, Wolf DR: **Base compositional structure of genomes.** *Genomics* 1992, **13:**1056–1064.

12. Prabhu VV: **Symmetry observations in long nucleotide sequences.** *Nucleic Acids Res* 1993, **21:**2797–2800.

13. Qi D, Cuticchia AJ: **Compositional symmetries in complete genomes.** *Bioinformatics* 2001, **17:**557–559.

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

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

16. Kozhukhin CG, Pevzner PA: **Genome inhomogeneity is determined mainly by WW and SS dinucleotides.** *Comput Appl Biosci* 1991, **7:**39–49.

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

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

19. Sanchez J, Jose MV: **Analysis of bilateral inverse symmetry in whole bacterial**

chromosomes. *Biochem Biophys Res Commun* 2002, **299:**126–134.

Additional Files

File name: Additional File 1; File format: MS Excel file; Title: Table 2. Dinucleotide frequencies across archaeal and bacterial genomes (130 species); Description: Original data and statistical analysis.

Figures

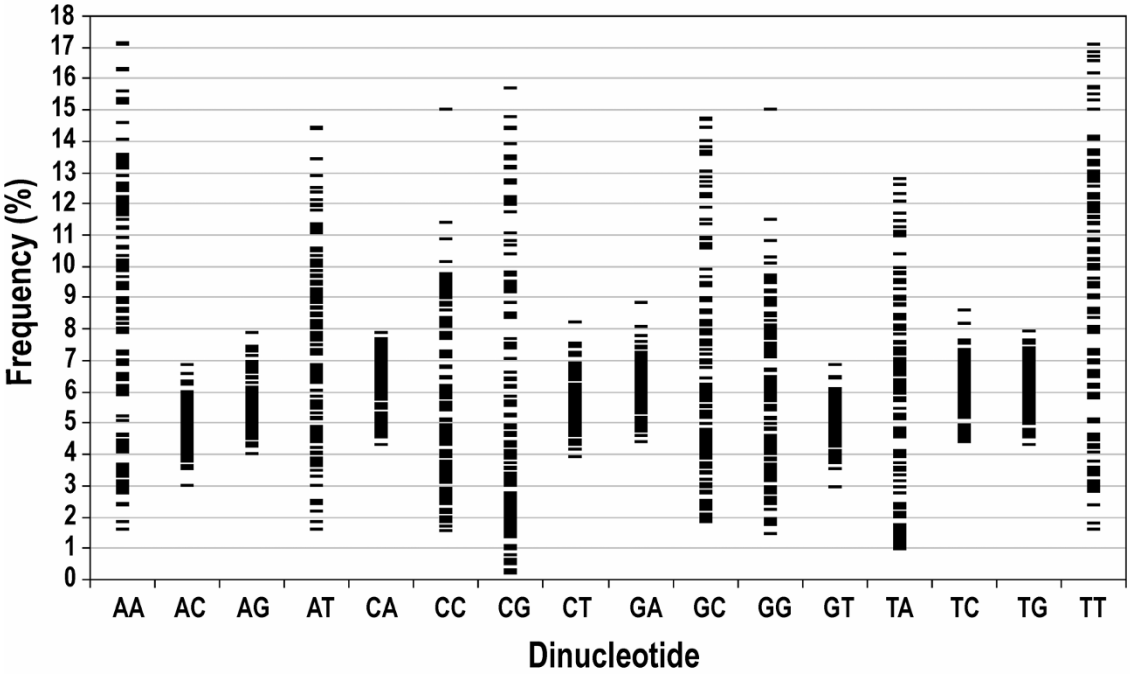

Figure Legends

Figure 1. Dinucleotide frequency distribution pattern of genomes of 130 species of archaea and bacteria. Each species is represented by a dash.

Tables

Table 1. Statistical analysis of dinucleotide frequencies and counts across 130 genomes

Dinucleotide	Mean (%)	Min[a] (%)	Max[b] (%)	s[c]	CV[d] (%)	r[e]	Slope[f]	Intercept[f]	P (χ^2)[g] (counts/kb)
AA	8.85	1.58	17.15	4.06	45.90	0.508	0.612	132698.03	0
AC	4.93	2.98	6.87	0.66	13.42	0.983	0.884	14391.96	0.803
AG	5.60	4.02	7.87	0.78	14.01	0.979	1.078	-8979.26	0.187
AT	7.46	1.59	14.42	2.85	38.20	0.670	0.811	65348.50	4.883×10^{-213}
CA	6.14	4.32	7.85	0.77	12.53	0.976	0.921	11480.38	0.606
CC	5.59	1.52	15.02	2.56	45.74	0.884	0.513	73343.12	1.471×10^{-233}
CG	5.82	0.21	15.68	4.15	71.27	0.846	0.348	102514.23	0
CT	5.60	3.92	8.21	0.80	14.33	0.978	1.075	-8664.17	0.113
GA	6.00	4.39	8.84	0.82	13.59	0.980	0.901	15770.75	0.184
GC	6.64	1.82	14.70	3.55	53.47	0.869	0.436	100284.22	0
GG	5.59	1.47	15.00	2.56	45.75	0.883	0.513	73441.69	1.210×10^{-234}
GT	4.93	2.93	6.85	0.66	13.44	0.982	0.886	14128.59	0.758
TA	5.84	0.97	12.81	3.34	57.13	0.321	0.345	131223.80	0
TC	6.00	4.37	8.59	0.81	13.52	0.980	0.900	15640.34	0.224
TG	6.15	4.30	7.94	0.77	12.56	0.975	0.924	11153.34	0.613
TT	8.86	1.59	17.09	4.07	45.97	0.504	0.603	134771.77	0

[a] minimum; [b] maximum; [c] standard deviation; [d] coefficient of variation; [e] correlation coefficient for the relationship between the observed counts and the expected counts; [f] slope or intercept of the best-fitted line for the observed counts vs. the expected counts; [g] P value for the χ^2 calculated for the difference between the observed values and the expected values (average counts/kb).