DNA Research **16**, 325–343, (2009) doi:10.1093/dnares/dsp021

A Study in Entire Chromosomes of Violations of the Intra-strand Parity of Complementary Nucleotides (Chargaff's Second Parity Rule)

B.R. Powdel[1], Siddhartha Sankar Satapathy[2], Aditya Kumar[3,†], Pankaj Kumar Jha[3,‡], Alak Kumar Buragohain[3], Munindra Borah[1], and Suvendra Kumar Ray[3,*]

Department of Mathematical Sciences, Tezpur University, Tezpur, Assam 784 028, India[1]; Department of Computer Science and Engineering, Tezpur University, Tezpur, Assam 784 028, India[2] and Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam 784 028, India[3]

(Received 11 June 2009; accepted 24 September 2009; published online 27 October 2009)

Abstract

Chargaff's rule of intra-strand parity (ISP) between complementary mono/oligonucleotides in chromosomes is well established in the scientific literature. Although a large numbers of papers have been published citing works and discussions on ISP in the genomic era, scientists are yet to find all the factors responsible for such a universal phenomenon in the chromosomes. In the present work, we have tried to address the issue from a new perspective, which is a parallel feature to ISP. The compositional abundance values of mono/oligonucleotides were determined in all non-overlapping sub-chromosomal regions of specific size. Also the frequency distributions of the mono/oligonucleotides among the regions were compared using the Kolmogorov–Smirnov test. Interestingly, the frequency distributions between the complementary mono/oligonucleotides revealed statistical similarity, which we named as intra-strand frequency distribution parity (ISFDP). ISFDP was observed as a general feature in chromosomes of bacteria, archaea and eukaryotes. Violation of ISFDP was also observed in several chromosomes. Chromosomes of different strains belonging a species in bacteria/archaea (*Haemophilus influenza*, *Xylella fastidiosa* etc.) and chromosomes of a eukaryote are found to be different among each other with respect to ISFDP violation. ISFDP correlates weakly with ISP in chromosomes suggesting that the latter one is not entirely responsible for the former. Asymmetry of replication topography and composition of forward-encoded sequences between the strands in chromosomes are found to be insufficient to explain the ISFDP feature in all chromosomes. This suggests that multiple factors in chromosomes are responsible for establishing ISFDP.

Key words: chromosome; nucleotide composition; Chargaff's second parity rule; intra-strand frequency distribution parity; DNA replication

Introduction

Chargaff's first parity rule based on the nucleotide composition of double-stranded DNA states that the complementary nucleotides have the same abundance values.[1,2] This is explained by the DNA double-helix model in which A pairs only with T and G pairs only with C.[3] Chargaff and his colleagues[4,5] came with a similar observation of compositional relationship between the complementary nucleotides even within individual DNA strands of bacterial chromosomes. In the post-genomic era, this intra-strand relationship between the complementary nucleotides is observed in double-stranded genomes of viruses, bacteria, archaea and eukaryotes, which is known as Chargaff's second parity rule or intra-strand parity (ISP).[2] There is no such defined rule to

Edited by Hiroyuki Toh

* To whom correspondence should be addressed. Tel. +91 3712-267007/008/009. Fax. +91 3712-267005/6. E-mail: suven@tezu.ernet.in

† Present address: Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.

‡ Present address: MS-04/603, Kendriya Vihar, Sector 56, Gurgaon, Haryana 122011, India.

describe ISP in chromosomes like the base-pairing rule in Chargaff's first parity. ISP is also observed between the complementary oligonucleotides in chromosomes,[6–9] which has been attributed to genome-wide large-scale inversion, inversion transposition[10] and coding sequence compositional symmetry between the strands.[9] Violation of ISP is observed with respect to organellar (mitochondria and plastids) genomes of some organisms, single-stranded viral genomes or any RNA genome.[11–13]

Theoretically, under no strand bias in terms of mutation and selection, the base complementary relationship easily explains the presence of ISP in chromosomes.[14,15] However, several evidences now prove that both the strands are not identical in terms of mutation/selection.[16] This results into violation of ISP in sub-chromosomal regions. Longer the sub-chromosomal region, smaller is the violation of ISP observed.[17] The mechanisms that are responsible to cause violation are defined under three categories.[18] First, DNA replication: leading strand (LeS) is found to be composed of more K nucleotides (G and T) than the complementary M (A and C) nucleotides and the reverse holds true for the lagging strand (LaS).[19] This is due to the fact that the LeS which functions as the template for Okazaki fragment synthesis (functions as template for LaS) remains exposed more as single-stranded than the LaS (functions as template for LeS) during replication that results into higher deamination of the cytosine residues[20,21] in LeS (cytosine gets deaminated 140 times faster in ssDNA than in dsDNA[22]). In addition, the influence of Okazaki fragments and the sliding DNA clamp proteins associated with the synthesis of LaS create functional asymmetry of the mismatch repairing system on DNA.[23] Second, transcription: genes are preferentially located in the LeS than in the LaS to avoid head on collision between the machineries of replication and transcription.[24] During transcription, the non-template strand remains more exposed as single-stranded than the template strand, which causes asymmetry in cytosine deamination between the strands.[22] The transcription-coupled repair system also acts only upon the template strand and thereby contributes to the strand asymmetry.[25] Third, translation: uses of synonymous codons are influenced by differential abundance of tRNA molecules which results into the differential abundance of complementary nucleotides at the third position of family box codons. This causes parity violation.[14] In spite of these factors favoring violations of the parity in chromosomes, ISP is observed in an entire chromosome due to the cancellation effect of the local violations in opposite directions.[14]

Evolutionary biologists are more interested to understand the role of mutation and/or selection in the violation of ISP by analyzing the weakly selected or selectively neutral regions (third position of family box codons and non-coding regions) in chromosomes.[14,26] Whether any specific feature(s) is/are associated with chromosomes exhibiting ISP is yet to be understood. Shioiri and Takahata[27] studied ISP by finding out the total AT skew (ATS) and GC skew (GCS) in the chromosomes of several bacteria. In their study, out of 36 bacterial chromosomes, *Xylella fastidiosa* exhibited maximum ATS and GCS. They observed variable ATS/GCS among chromosomes of different strains of a species as well as chromosomes within a bacterial cell. They also observed ATS and GCS may be different from each other within a chromosome. Since, they did not do any statistical analysis of the skew, the significance of the variability observed among chromosomes was not discussed by them. The usual statistical tool used to find out ISP in chromosomes is a correlation analysis of oligonucleotides abundance described by Prabhu.[6] The ISP study between the complimentary mononucleotides is important because it has been proven that oligonucleotide parity and mononucleotide parity are independent.[8] Baisnée *et al.*[8] studied parity in chromosomes by measuring the S^1 index which is defined as the sum of the absolute values of the differences between complementary oligonucleotides (*n* mer) frequencies (*n* varies from 1 to 9 mer). Both these methods do not measure the statistical significance of differences between the abundance values of a mono/oligonucleotide and its reverse complement. For example, if a chromosome carries significant similarity between the abundance values of A and T but carries significant difference between the abundance values of G and C, this will not be identified separately. Similarly, the above methods are unable to find out parity violations in chromosomes with respect to the abundance values of an oligonucleotide and its reverse complement. We have developed a methodology here that can independently study ISP between S nucleotides (any oligonucleotide and its reverse complement) as well as between W nucleotides using the abundance values of mononucleotides. We use the well-known Kolmogorov–Smirnov (KS) test to study the frequency distribution of the compositional abundance values of the mononucleotides in a chromosome sequence, which gives the statistical significance of the similarity between the distributions of complementary nucleotides. This we called as intra-strand frequency distribution parity (ISFDP), which has been used here to study the chromosomes of bacteria, archaea and eukaryotes.

Materials and methods

Frequency distribution calculation

Chromosome sequences of different bacteria, archaea and eukaryotes (Tables 1–3) were obtained from the genome information broker, DDBJ site (www.gib.genes.nig.ac.jp). Bacterial chromosomes were chosen randomly from the database starting the genus name from A to Z. Chromosome sequences of different strains belonging to the same species in the case of bacteria were taken in several cases to do the intra-species comparison. Each chromosome sequence was divided into smaller-size sequences of 1000 nucleotides each starting from the beginning, and the abundance value of the four nucleotides was determined using the computer program (developed for this study). The distribution of the abundance values of complementary nucleotides in different fragments were analyzed by the KS non-parametric test using XLSTAT program[28–30] (Kovach Computing Services, Anglesey, Wales). H_0: distribution patterns of any two nucleotides/oligonucleotides in a chromosome are similar; H_A: there is a difference between the two distributions. Owing to the large sample size, similarity was considered at the *P*-value of >0.01, weak similarity was considered at the *P*-value between 0.01 and 10^{-4}, and the value of $<10^{-4}$ was considered as strong violation similarity. Group-frequency distributions of the abundance values were plotted to observe the frequency-distribution parity. In the case of the di- and trinucleotides, the abundance values were determined using a different computer program (developed here for this study) in the segments for the 16 dinucleotides and 64 trinucleotides. The analysis was done as described for the mononucleotides earlier.

Angular replication asymmetry of the chromosomes was calculated with the help of the information on *ori* (origin) and *ter* (termination) cited in the websites (http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/ and http://pbil.univ-lyon1.fr/software/Oriloc/oriloc.html). The chromosomal region starting from *ori* to *ter* was considered as the leading region in the Watson strand (Ws) and the remaining portion of the chromosome as the lagging region. For a circular chromosome, the angular replication asymmetry was calculated as the amount of angular distance of leading region deviating from 180°.

Proportionate distribution of forward- and reverse-encoded sequences in a DNA strand

From the DDBJ site, only coding sequences were downloaded. A continuous stretch of the nucleotide sequence was made from all the sequences by removing the gene names. This resembled a DNA strand only composed of forward-encoded sequences. Frequency distribution analysis was done on this. In another approach, 50% of the above strand was made reverse complement by *in silico* followed by joining with the rest. This resembled a DNA strand composed of 50% forward-encoded and 50% reverse-encoded sequences. Frequency-distribution study was carried out as described above.

Identification of leading and LaS region

ATS and GCS analyses of the chromosome sequences were done as described earlier.[21] This was used to find out the tentative leading and lagging portions in a DNA strand.

Relative proportion of coding sequence distribution

This was found out by deducting ORF numbers between Ws (top strand) and Crick strand (Cs: bottom strand) followed by dividing that with the total number of ORFs. Gene orientation information was obtained from the website (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi).

Results

ISFDP in chromosomes of bacteria

In this study, a total of 112 bacterial chromosomes were considered, which includes different lineages of bacteria such as protobacteria, cyanobacteria, firmicutes, actinobacteria etc. Samples from each group were taken randomly. The bacteria included in the sample comprised a GC% variation from a minimum of 28% to a maximum of 75% and chromosome size variation from 580 kb to a maximum of 9105 kb. We have studied the frequency distributions of the abundance values of mononucleotides in the uniform sub-chromosomal length of 1000 nucleotides. A collective analysis of the nucleotide abundance values from all the segments of a chromosome was done by frequency distribution smooth curves using Microsoft Excel, and the similarity of the distributions of two complementary nucleotides was tested using the KS test (XL-Stat; http://www.xlstat.com/en/download). Figure 1A(i), B(i), C(i), D(i) and E(i) represents the smooth curves of frequency distributions of nucleotides in chromosomes *Campylobacter jejuni* RM1221 (30.31%), *Escherichia coli* K12 MG1655 (50.79%), *Xanthomonas campestris* pv. *campestris* (*Xcc*; 65.07%), *X. fastidiosa* 9a5c (52.68%) and *X. fastidiosa* Temecula (51.78%). Smooth curves of complementary nucleotides overlap with each other in the first three chromosomes, whereas those of non-complementary ones do not. In the fourth chromosome, none of the curves overlap with each other. In *E. coli*

Table 1. ISFDP analysis in bacterial chromosomes

Serial number	Strain name	Size (kb)	GC%	KS (W)	KS (S)	$\lvert(\sum A - \sum T)\rvert / (\sum A + \sum T)$	$\lvert(\sum G - \sum C)\rvert / (\sum G + \sum C)$	Bacterial group	TB (°)
1	*Acinetobacter* sp. ADP1	3598	40.43	0.745	*0.006*	0.00068	0.00484	G-Proteobacteria	7.07
2	*Actinobacillus pleuropneumoniae* L20 serotype 5b	2274	41.3	0.436	0.819	0.00187	0.00109		NA
3	*Actinobacillus succinogenes* 130Z	2319	44.91	0.312	0.291	0.00232	0.00291		
4	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	4744	61.55	0.88	0.19	0.00141	0.00139		
5	*Aeromonas salmonicida* subsp. *salmonicida* A449	4702	58.51	0.04	0.959	0.00215	0.00073		
6	*Agrobacterium tumefaciens* C58 (circular chromosome)	2841	59.38	**<0.0001**	**<0.0001**	0.00694	0.00967	A-Proteobacteria	7.37
7	*Alkaliphilus oremlandii* OhILAs	3123	36.26	**<0.0001**	**<0.0001**	0.00615	0.01324	Firmicutes	NA
8	*Anaeromyxobacter dehalogenans* 2CP-C	5013	74.9	0.077	*0.001*	0.00476	0.00249	D-Proteobacteria	70.57
9	*Anaeromyxobacter* sp. Fw109-5	5277	73.53	0.712	*0.008*	0.00073	0.00216		7.48
10	*Bacillus anthracis* Ames	5227	35.38	*0.004*	**<0.0001**	0.00215	0.00581	Firmicutes	NA
11	*Bacillus anthracis* 'Ames Ancestor'	5227	35.38	*0.003*	**<0.0001**	0.00215	0.00582		7.48
12	*Bacillus anthracis* Sterne	5228	35.38	*0.008*	**<0.0001**	0.00221	0.00588		7.46
13	*Bacillus subtilis*	4214	43.52	0.219	0.234	0.00212	0.00224		13.69
14	*Bacillus thuringiensis* Al Hakam	5257	35.43	0.123	*0.002*	0.00042	0.00081		NA
15	*Bacillus thuringiensis* serovar konkukian 97-27	5237	35.41	0.015	**<0.0001**	0.00194	0.00438		3.98
16	*Bordetella parapertussis* 12822	4773	68.1	0.433	**<0.0001**	0.00247	0.00776	B-Proteobacteria	37.01
17	*Bordetella pertussis* Tohama 1	4086	67.72	0.861	**<0.0001**	0.00022	0.00390		71.28
18	*Bradyrhizobium japonicum* USDA 110	9105	64.06	0.512	0.31	0.00070	0.00038	A-Proteobacteria	7.07
19	*Bradyrhizobium* sp. BTAi1	8264	64.92	0.381	0.01	0.00100	0.00163		NA
20	*Brucella melitensis* 16M	1177	57.35	0.472	*0.008*	0.00227	0.00312		
21	*Campylobacter concisus* 13826	2052	39.43	0.033	0.048	0.00038	0.00599	E-Proteobacteria	
22	*Campylobacter curvus* 525.92	1971	44.54	0.028	0.752	0.00745	0.00282		
23	*Campylobacter jejuni* RM1221	1777	30.31	0.574	0.23	0.00330	0.00436		8.69
24	*Campylobacter jejuni* subsp. *jejuni* 81116	1628	30.54	0.491	0.029	0.00250	0.00613		NA
25	*Campylobacter jejuni* subsp. *jejuni* NCTC 11168	1641	30.55	0.067	0.132	0.00296	0.00457		10.25
26	Candidatus *Desulfococcus oleovorans* Hxd3	3944	56.17	0.258	0.133	0.00199	0.00157	Firmicutes	NA
27	*Caulobacter crescentus* CB15	4016	67.22	0.042	0.171	0.00396	0.00188	A-Proteobacteria	8.56
28	*Chlamydia muridarum* Nigg	1072	40.34	0.221	0.853	0.00107	0.00337	Chlamydiae	1.17
29	*Chlamydia trachomatis* AHAR-13	1044	41.31	0.228	0.284	0.00230	0.00059		1.30
30	*Chlamydophila abortus* S263	1144	39.87	0.534	*0.002*	0.00065	0.00361		0.57
31	*Coxiella burnetii* Dugway 7E9-12	2158	42.44	*0.004*	*0.001*	0.00592	0.00573	G-Proteobacteria	NA
32	*Coxiella burnetii* RSA 493	1995	42.66	0.014	0.467	0.00198	0.00029		31.15
33	*Desulfovibrio desulfuricans* G20	3730	57.84	0.59	*0.001*	0.00189	0.00322	Firmicutes	10.70

34	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	3462	63.01	0.3	0.159	0.00152	0.00106	D-Proteobacteria	NA
35	*Desulfovibrio vulgaris* subsp. *vulgaris* Hildenborough	3570	63.14	0.557	0.082	0.00143	0.00024		4.78
36	*Enterobacter sakazakii* ATCC BAA-894	4368	56.77	0.167	0.388	0.00359	0.00044	G-Proteobacteria	NA
37	*Enterobacter* sp. 638	4518	52.98	0.645	0.39	0.00169	0.00163		NA
38	*Escherichia coli* 536	4938	50.52	0.714	0.084	0.00062	0.00328		7.40
39	*Escherichia coli* APEC O1	5082	50.55	0.779	0.576	0.00032	0.00070		NA
40	*Escherichia coli* CFT073	5231	50.48	0.112	0.92	0.00173	0.00080		5.66
41	*Escherichia coli* E24377A	4979	50.62	0.736	0.128	0.00205	0.00212		NA
42	*Escherichia coli* HS	4643	50.82	0.328	0.469	0.00151	0.00207		
43	*Escherichia coli* K12 MG1655	4639	50.79	0.732	0.587	0.00054	0.00113		4.28
44	*Escherichia coli* UTI89	5065	50.6	0.51	0.237	0.00076	0.00203		3.70
45	*Escherichia coli* W3110	4646	50.8	0.873	0.729	0.00073	0.00091		12.64
46	*Frankia alni* ACN14A chromosome	7497	72.82	0.463	0.036	0.00141	0.00139	Actinobacteria	NA
47	*Frankia* sp. CcI3	5433	70.08	0.808	0.662	0.00129	0.00017		
48	*Haemophilus influenzae* 86-028NP	1914	38.16	0.886	0.654	0.00089	0.00044	G-Proteobacteria	
49	*Haemophilus influenzae* PittEE	1813	38.04	0.544	0.038	0.00054	0.00317		
50	*Haemophilus influenzae* PittGG	1887	38.01	0.125	**<0.0001**	0.00005	0.01016		
51	*Haemophilus influenzae* Rd KW20	1830	38.15	0.154	*0.004*	0.00298	0.00472		46.61
52	*Helicobacter acinonychis* Sheeba	1553	38.18	*0*	0.596	0.00869	0.00164	E-Proteobacteria	NA
53	*Helicobacter hepaticus* ATCC 51449	1799	35.93	0.161	**<0.0001**	0.00499	0.01518		46.54
54	*Helicobacter pylori* J99	1643	39.19	0.246	0.256	0.00259	0.00510		10.97
55	*Lactobacillus acidophilus* NCFM	1993	34.72	0.382	**<0.0001**	0.00066	0.01644	Firmicutes	19.54
56	*Lactobacillus brevis* ATCC 367	2291	46.22	0.023	**<0.0001**	0.00271	0.02882		NA
57	*Lactobacillus delbrueckii* subsp. *bulgaricus* ATCC BAA-365	1856	49.69	0.491	0.264	0.00201	0.00087		
58	*Lactobacillus reuteri* F275	1999	38.87	*0.001*	**<0.0001**	0.00122	0.01040		
59	*Lactococcus lactis* subsp. *cremoris* MG1363	2529	35.75	0.233	0.056	0.00352	0.00524		
60	*Lactococcus lactis* subsp. *cremoris* SK11	2438	35.86	0.399	0.521	0.00147	0.00136		
61	*Magnetococcus* sp. MC-1	4719	54.17	*0.001*	**<0.0001**	0.00490	0.01198	Magnetococcus	
62	*Magnetospirillum magneticum* AMB-1	4967	65.09	0.031	**<0.0001**	0.00339	0.00288	A-Proteobacteria	2.14
63	*Methylobacillus flagellatus* KT	2971	55.72	0.03	0.916	0.00226	0.00135	B-Proteobacteria	10.57
64	*Methylococcus capsulatus* Bath	3304	63.59	0.145	*0.004*	0.00150	0.00287	G-Proteobacteria	NA
65	*Mycobacterium leprae* TN	3268	57.8	*0.003*	**<0.0001**	0.00378	0.00609	Actinobacteria	7.04
66	*Mycobacterium* sp. KMS	5737	68.44	0.389	0.478	0.00030	0.00060		NA
67	*Mycobacterium tuberculosis* F11	4424	65.62	0.366	*0.007*	0.00006	0.00198		
68	*Mycobacterium ulcerans* Agy99	5631	65.47	**<0.0001**	**<0.0001**	0.00433	0.00374		

Continued

Table 1. Continued

Serial number	Strain name	Size (kb)	GC%	KS (W)	KS (S)	$\lvert(\sum A - \sum T)\rvert/(\sum A + \sum T)$	$\lvert(\sum G - \sum C)\rvert/(\sum G + \sum C)$	Bacterial group	TB (°)
69	*Mycoplasma gallisepticum* R	996	31.45	0.18	0.615	0.00626	0.00021	Tenericutes	9.32
70	*Mycoplasma genitalium* G37	580	31.69	*0*	0.148	0.01219	0.00433		3.75
71	*Mycoplasma hyopneumoniae* J	897	28.52	0.033	0.599	0.01020	0.00067		NA
72	*Mycoplasma pneumoniae* M129	816	40.01	*0.001*	0.115	0.01767	0.00243		16.23
73	*Neisseria gonorrhoeae* FA 1090	2153	52.69	0.07	0.033	0.00601	0.00144	B-Proteobacteria	9.20
74	*Neisseria meningitidis* MC58	2273	51.52	0.695	*0.004*	0.00135	0.00806		NA
75	*Nitrobacter hamburgensis* X14	4406	61.72	0.332	0.53	0.00112	0.00041	A-Proteobacteria	
76	*Nitrobacter winogradskyi* Nb-255	3402	62.05	0.011	**<0.0001**	0.00323	0.00294		37.15
77	*Nitrosococcus oceani* ATCC 19707	3481	50.32	0.02	0.056	0.00530	0.00243	G-Proteobacteria	8.39
78	*Nitrosomonas eutropha* C91	2661	48.49	0.992	0.318	0.00043	0.00162	B-Proteobacteria	NA
79	*Nostoc* sp. PCC 7120	6413	41.35	0.134	0.857	0.00129	0.00162	Cyanobacteria	
80	*Pseudomonas entomophila* L48 chromosome	5888	64.16	0.657	0.251	0.00078	0.00173	G-Proteobacteria	1.99
81	*Pseudomonas fluorescens* PfO-1	6438	60.52	*0.003*	0.028	0.00443	0.00222		3.18
82	*Pseudomonas putida* F1	5959	61.86	0.602	0.013	0.00113	0.00187		36.81
83	*Ralstonia eutropha* H16	2912	66.78	0.238	0.47	0.00483	0.00023	B-Proteobacteria	NA
84	*Ralstonia solanacearum* GMI1000 chromosome	3716	67.04	0.056	**<0.0001**	0.00636	0.00581		22.40
85	*Rhizobium etli* CFN 42	4381	61.27	0.107	**<0.0001**	0.00175	0.01177	A-Proteobacteria	17.65
86	*Rhizobium leguminosarum* bv. *viciae* 3841	5057	61.09	*0.001*	**<0.0001**	0.00363	0.01196		NA
87	*Rickettsia bellii* RML369-C	1522	31.65	*0*	**<0.0001**	0.00859	0.01514		26.08
88	*Rickettsia conorii* Malish 7	1268	32.44	0.584	0.052	0.00294	0.00634		16.28
89	*Rickettsia rickettsii* 'Sheila Smith'	1257	32.47	0.575	*0.002*	0.00182	0.00767		NA
90	*Rickettsia typhi* Wilmington	1111	28.92	0.919	*0.007*	0.00020	0.01395		26.15
91	*Salmonella enterica* subsp. *enterica* serovar Typhi CT18	4809	52.09	0.267	0.043	0.00151	0.00152	G-Proteobacteria	9.85
92	*Salmonella typhimurium* LT2	4857	52.22	0.89	0.585	0.00043	0.00008		3.58
93	*Shigella boydii* Sb227	4519	51.21	0.571	*0.001*	0.00022	0.00249		11.05
94	*Shigella flexneri* 58401	4574	50.92	0.48	0.268	0.00147	0.00214		NA
95	*Staphylococcus aureus* RF122	2742	32.78	0.788	0.427	0.00130	0.00247	Firmicutes	0.10
96	*Staphylococcus epidermidis* ATCC 12228	2499	32.1	**<0.0001**	**<0.0001**	0.01246	0.01087		21.12
97	*Staphylococcus haemolyticus* JCSC1435	2685	32.79	*0.001*	*0*	0.00584	0.00643		NA
98	*Streptococcus mutans* UA159	2030	36.83	0.111	0.046	0.00403	0.00679		
99	*Streptococcus pyogenes* MGAS2096	1860	38.73	0.619	0.15	0.00133	0.00154		3.71
100	*Streptococcus thermophilus* CNRZ1066	1796	39.08	0.05	0.863	0.00537	0.00459		2.63
101	*Streptomyces coelicolor* A3(2)	8667	72.12	*0.001*	0.037	0.00394	0.00134	Actinobacteria	NA

102	*Thermotoga maritima* MSB8	1860	46.25	0.171	**<0.0001**	0.00344	0.01548	Thermotogae	39.15
103	*Thermotoga petrophila* RKU-1	1824	46.09	0.733	**<0.0001**	0.00013	0.01687		NA
104	*Thiobacillus denitrificans* ATCC 25259	2909	66.07	0.962	0.086	0.00027	0.00059	B-Proteobacteria	5.70
105	*Vibrio cholerae* O395	3024	47.78	**<0.0001**	0.069	0.00514	0.00105	G-Proteobacteria	NA
106	*Vibrio fischeri* ES114	1332	37.03	*0.001*	0.037	0.00994	0.00491		
107	*Xanthomonas campestris* pv. *campestris* ATCC 33913	5076	65.07	0.196	0.719	0.00302	0.00038		
108	*Xanthomonas oryzae* pv. *oryzae* KACC 10331	4941	63.69	0.87	0.499	0.00104	0.00065		
109	*Xylella fastidiosa* 9a5c	2679	52.68	**<0.0001**	**<0.0001**	0.04727	0.05291		62.97
110	*Xylella fastidiosa* Temecula 1	2519	51.78	0.044	*0*	0.00379	0.01093		6.44
111	*Yersinia pestis* CO92	4653	47.64	0.649	*0.001*	0.00090	0.00520		NA
112	*Yersinia pseudotuberculosis* IP32953	4744	47.61	0.969	*0.001*	0.00124	0.00496		

TB, termination bias. Chromosomes of bacteria analyzed in this study. The KS test for significance between the frequency distribution of complementary nucleotide values are given as: KS (W) between A and T and KS (S) between G and C. In bacteria, archaea and eukaryotes, *P*-values of $<10^{-4}$ (strong violation of ISFDP) are shown in bold and *P*-values of <0.01 but$\geq 10^{-4}$ (weak violation of ISFDP) are shown in italics. The *P*-value between 10^{-4} and 10^{-3} is shown as 0.000. Relative absolute abundance value difference between the complementary nucleotides is given by $|(\sum A - \sum T)|/(\sum A + \sum T)$ and $|(\sum G - \sum C)|/(\sum G + \sum C)$ for ATS and GCS, respectively. In chromosome of *X. fastidiosa* 9a5c, the GCS/ATS value is highest suggesting that the difference between the abundance values of complementary nucleotides is high. The *P*-value by the KS test is in concordant with the ATS/GCS suggesting that the abundance difference can be represented by the frequency distribution study of the nucleotides. Similar relation is also observed in other chromosomes.

Table 2. ISFDP analysis in archaea chromosomes

Serial number	Strain name	Size (kb)	GC%	KS (W)	KS (S)	$\lvert(\sum A - \sum T)\rvert / (\sum A + \sum T)$	$\lvert(\sum G - \sum C)\rvert / (\sum G + \sum C)$	Archaea group
1	*Aeropyrum pernix* K1	1670	56.3	*0.001*	0.025	0.01292	0.00695	Crenarchaeota
2	*Archaeoglobus fulgidus* DSM 4304	2179	48.5	0.037	0.093	0.00365	0.00350	Euryarchaeota
3	*Caldivirga maquilingensis* IC-167	2078	43.08	0.586	0.643	0.00146	0.00104	Crenarchaeota
4	Candidatus *Methanoregula boonei* 6A8	2543	54.51	0.058	0.191	0.00311	0.00108	Euryarchaeota
5	*Cenarchaeum symbiosum*	2046	57.34	0.101	*0.006*	0.00574	0.00161	Crenarchaeota
6	*Haloarcula marismortui* ATCC 43049 chromosome 1	3132	62.35	0.252	0.905	0.01075	0.00024	Euryarchaeota
7	*Halobacterium* sp. NRC-1	2015	67.88	0.862	0.313	0.00056	0.00151	
8	*Haloquadratum walsbyi* DSM 16790	3133	47.85	0.578	0.027	0.00160	0.00523	
9	*Hyperthermus butylicus* DSM 5456	1668	53.7	0.019	0.908	0.00531	0.00100	Crenarchaeota
10	*Ignicoccus hospitalis* KIN4/I	1298	56.5	0.118	0.901	0.00199	0.00014	
11	*Metallosphaera sedula* DSM 5348	2192	46.21	*0*	**<0.0001**	0.00668	0.01423	Crenarchaeota
12	*Methanobrevibacter smithii* ATCC 35061	1854	31.02	**<0.0001**	**<0.0001**	0.02048	0.03768	Euryarchaeota
13	*Methanocaldococcus jannaschii* DSM 2661	1666	31.4	0.132	0.031	0.00450	0.01128	
14	*Methanococcoides burtonii* DSM 6242	2576	40.74	0.078	*0.002*	0.00266	0.00845	
15	*Methanococcus aeolicus* Nankai-3	1570	30.02	0.218	0.52	0.00399	0.00063	
16	*Methanococcus maripaludis* C5	1781	32.99	*0.001*	0.065	0.00846	0.00454	
17	*Methanococcus maripaludis* C6	1745	33.4	0.045	0.045	0.00553	0.00224	
18	*Methanococcus maripaludis* C7	1773	33.27	0.256	0.784	0.00430	0.00088	
19	*Methanococcus maripaludis* S2	1662	33.08	0.021	0.08	0.00619	0.00259	
20	*Methanococcus vannielii* SB	1721	31.31	0.505	0.519	0.00364	0.00400	
21	*Methanocorpusculum labreanum* Z	1806	49.97	0.606	0.05	0.00097	0.00404	
22	*Methanoculleus marisnigri* JR1	2479	62.04	0.816	0.745	0.00234	0.00000	
23	*Methanopyrus kandleri* AV19	1696	61.22	0.556	0.032	0.00230	0.00471	
24	*Methanosaeta thermophila* PT	1880	53.53	0.673	*0.004*	0.00018	0.00595	
25	*Methanosarcina acetivorans* C2A	5752	42.67	**<0.0001**	0.839	0.00628	0.00083	
26	*Methanosarcina barkeri* Fusaro	4838	39.27	**<0.0001**	*0.003*	0.00475	0.00391	
27	*Methanosarcina mazei* Goe1	4097	41.47	0.252	0.812	0.00212	0.00079	
28	*Methanosphaera stadtmanae* DSM 3091	1768	27.62	*0.002*	0.275	0.00897	0.00652	
29	*Methanospirillum hungatei* JF-1	3545	45.14	**<0.0001**	0.015	0.00951	0.00411	
30	*Methanothermobacter thermautotrophicus* Delta H	1752	49.52	0.022	0.114	0.00566	0.00166	
31	*Nanoarchaeum equitans* Kin4-M	491	31.55	0.549	0.177	0.00000	0.00127	Nanoarchaeota
32	*Natronomonas pharaonis* DSM 2160	2596	63.42	0.473	0.228	0.00174	0.00091	Euryarchaeota
33	*Nitrosopumilus maritimus* SCM1	1646	31.15	**<0.0001**	*0.002*	0.00921	0.00855	Crenarchaeota
34	*Picrophilus torridus* DSM 9790	1546	35.96	0.296	*0.008*	0.00096	0.00793	Euryarchaeota

35	*Pyrobaculum aerophilum* IM2	2223	51.34	*0.001*	**<0.0001**	0.00727	0.01022	Crenarchaeota
36	*Pyrobaculum arsenaticum* DSM 13514	2122	54.98	0.795	0.431	0.00138	0.00316	
37	*Pyrobaculum calidifontis* JCM 11548	2010	57.13	0.148	0.337	0.00294	0.00008	
38	*Pyrobaculum islandicum* DSM 4184	1827	49.58	0.305	0.436	0.00085	0.00183	
39	*Pyrococcus abyssi*	1766	44.69	0.652	0.574	0.00219	0.00342	Euryarchaeota
40	*Pyrococcus furiosus* DSM 3638	1909	40.75	0.754	0.757	0.00004	0.00094	
41	*Pyrococcus horikoshii* OT3	1739	41.86	0.133	*0.002*	0.00229	0.01262	
42	*Staphylothermus marinus* F1	1571	35.71	**<0.0001**	**<0.0001**	0.02078	0.02726	Crenarchaeota
43	*Sulfolobus acidocaldarius* DSM 639	2227	36.69	0.413	0.526	0.00309	0.00124	
44	*Sulfolobus solfataricus* P2	2993	35.77	*0.007*	0.747	0.00533	0.00241	
45	*Sulfolobus tokodaii 7*	2695	32.78	*0.005*	0.029	0.00521	0.00659	
46	*Thermococcus kodakarensis* KOD1	2089	51.98	0.062	0.328	0.00418	0.00160	Euryarchaeota
47	*Thermofilum pendens* Hrk 5	1782	57.66	0.014	*0.005*	0.00346	0.00665	Crenarchaeota
48	*Thermoplasma acidophilum* DSM 1728	1565	45.99	0.016	0.016	0.00680	0.00383	Euryarchaeota
49	*Thermoplasma volcanium* GSS1	1585	39.91	0.055	0.361	0.00404	0.00263	

Chromosomes of archaea analyzed in this study. The KS test for significance between the frequency distribution of complementary nucleotide values are given as KS (W) between A and T and KS (S) between G and C. In bacteria, archaea and eukaryotes, *P*-values of $<10^{-4}$ (strong violation of ISFDP) are shown in bold and *P*-values of <0.01 but $\geq 10^{-4}$ (weak violation of ISFDP) are shown in italics. The *P*-value between 10^{-4} and 10^{-3} is shown as 0.000. Relative absolute abundance value difference between the complementary nucleotides is given by $|(\sum A - \sum T)|/(\sum A + \sum T)$and $|(\sum G - \sum C)|/(\sum G + \sum C)$ for ATS and GCS, respectively. In chromosome of *X. fastidiosa* 9a5c, the GCS/ATS value is highest suggesting the difference between the abundance values of complementary nucleotides is high. The *P*-value by the KS test is in concordant with the ATS/GCS suggesting that the abundance difference can be represented by the frequency distribution study of the nucleotides. Similar relation is also observed in other chromosomes.

Table 3. ISFDP analysis in eukaryotes chromosomes

Serial number	Strain name	Size (kb)	GC%	KS (W)	KS (S)	$\lvert(\sum A - \sum T)\rvert/(\sum A + \sum T)$	$\lvert(\sum G - \sum C)\rvert/(\sum G + \sum C)$	Eukaryotes group
1	*Guillardia theta* nucleomorph chromosome 01	197	25.64	0.411	0.468	0.00080	0.00517	Cryptophyta
2	*Guillardia theta* nucleomorph chromosome 02	181	26.7	0.435	0.35	0.00451	0.00356	
3	*Guillardia theta* nucleomorph chromosome 03	175	26.81	0.671	0.403	0.00051	0.00622	
4	*Leishmania major* Friedlin chromosome 01	270	62.84	0.055	**<0.0001**	0.01290	0.02500	Euglenozoa
5	*Plasmodium falciparum* 3D7 chromosome 01	644	20.52	*0.001*	0.69	0.02184	0.01210	Alveolata
6	*Plasmodium falciparum* 3D7 chromosome 05	1344	19.32	*0.006*	*0.005*	0.01288	0.01482	
7	*Plasmodium falciparum* 3D7 chromosome 11	2036	18.95	0.043	0.027	0.00339	0.00994	
8	*Plasmodium falciparum* 3D7 chromosome 12	2272	19.31	0.05	0.677	0.00597	0.00376	
9	*Plasmodium falciparum* 3D7 chromosome 13	2733	19.11	0.105	0.266	0.00422	0.00914	
10	*Plasmodium falciparum* 3D7 chromosome 14	3292	18.43	0.258	0.062	0.00275	0.00730	
11	*Saccharomyces cerevisiae* S288C chromosome 01	231	39.14	0.731	0.088	0.00100	0.01231	Fungi
12	*Saccharomyces cerevisiae* S288C chromosome 04	1532	37.9	0.807	0.379	0.00240	0.00345	
13	*Saccharomyces cerevisiae* S288C chromosome 07	1091	38.05	0.285	0.85	0.00136	0.00080	
14	*Saccharomyces cerevisiae* S288C chromosome 12	1079	38.44	0.055	0.461	0.00325	0.00173	
15	*Saccharomyces cerevisiae* S288C chromosome 15	1092	38.13	0.181	0.64	0.00584	0.00387	
16	*Schizosaccharomyces pombe* 972h chromosome 01	5574	36.09	0.4	0.076	0.00073	0.00086	
17	*Schizosaccharomyces pombe* 972h chromosome 02	4510	35.92	0.461	0.825	0.00207	0.00039	
18	*Schizosaccharomyces pombe* 972h chromosome 03	2453	36.23	0.152	0.012	0.00217	0.00369	

Chromosomes of eukaryotes analyzed in this study. The KS test for significance between the frequency distribution of complementary nucleotide values are given as KS (W) between A and T and KS (S) between G and C. In bacteria, archaea and eukaryotes, *P*-values of $<10^{-4}$ (strong violation of ISFDP) are shown in bold and *P*-values of <0.01 but $\geq 10^{-4}$ (weak violation of ISFDP) are shown in italics. The *P*-value between 10^{-4} and 10^{-3} is shown as 0.000. Relative absolute abundance value difference between the complementary nucleotides is given by $|(\sum A - \sum T)|/(\sum A + \sum T)$ and $|(\sum G - \sum C)|/(\sum G + \sum C)$ for ATS and GCS, respectively. In chromosome of *X. fastidiosa* 9a5c, the GCS/ATS value is highest suggesting that the difference between the abundance values of complementary nucleotides is high. The *P*-value by the KS test is in concordant with the ATS/GCS suggesting that the abundance difference can be represented by the frequency distribution study of the nucleotides. Similar relation is also observed in other chromosomes.

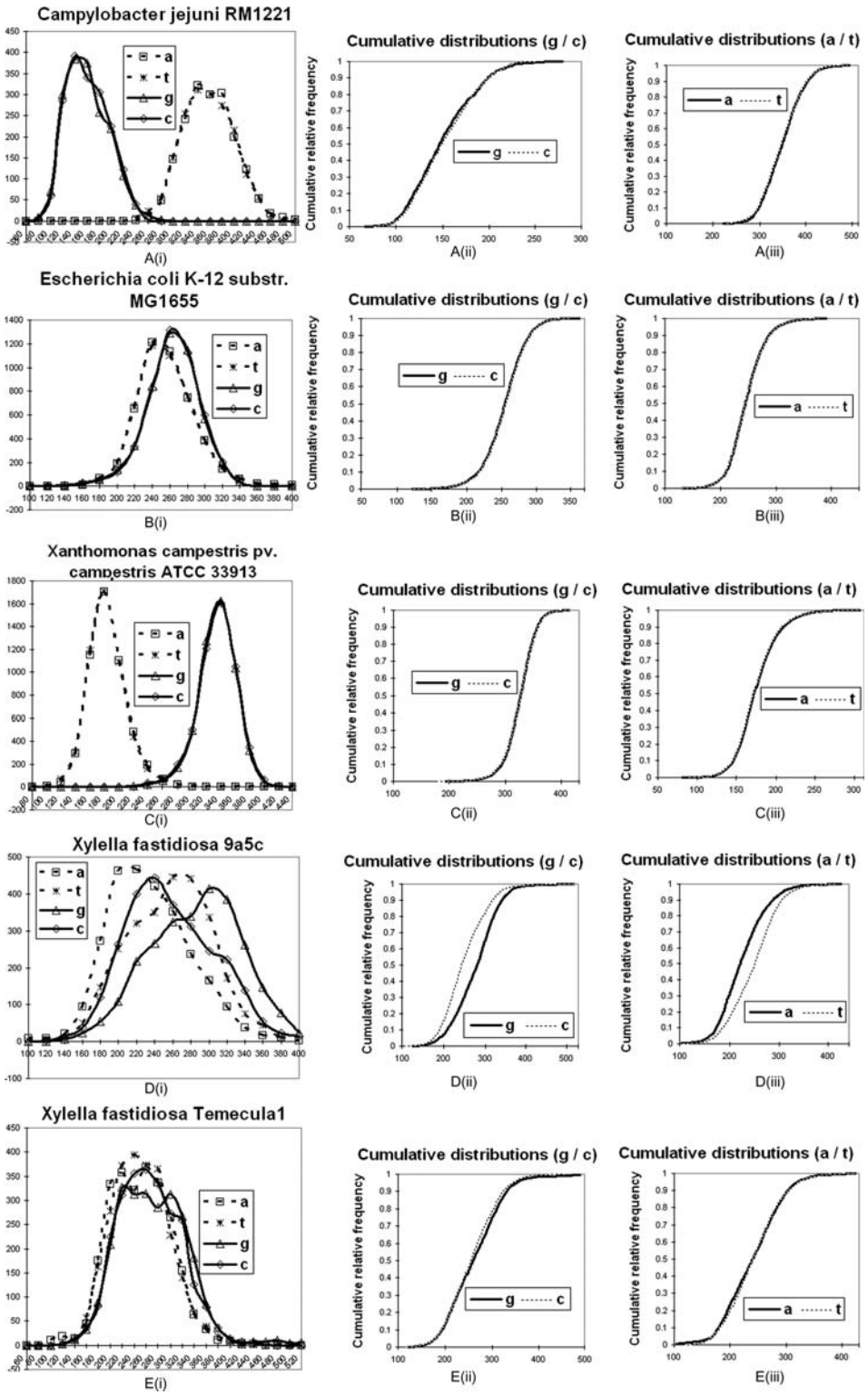

Figure 1. (**A**–**E**) Frequency distribution of nucleotides in chromosomes. Smooth curves present the group-frequency distribution of the four nucleotides a (square), t (asterisk), g (triangle) and c (rhombus). The *X*-axis represents the abundance values of the nucleotide spanning a range, whereas the *Y*-axis represents the frequency of the abundance values. In (A), the chromosome is AT rich; in (B), the chromosome is composed of similar AT and GC and in (C), the chromosome is GC rich. This is also evident from the group-frequency distribution curve. The smooth frequency curves of complementary nucleotides in these chromosomes are overlapping with each other. The KS test is shown for S and W nucleotides separately adjacent to the figures, respectively [a(ii, iii)–e(ii, iii)]. The KS test is in concordance with the curve obtained by smoothing group-frequency distribution. In (D) and (E), the group-frequency distribution for the chromosomes of two strains of *X. fastidiosa* is shown. In 9a5c strain chromosome, the smooth frequency curve between the complementary nucleotides does not overlap which is also suggested by the KS test. However, in Temecula 1 strain chromosome, the parity is maintained.

chromosome [Fig. 1B(i)], all the four smooth frequency curves are close to each other due to the closeness of the abundance values of the nucleotides, whereas in the graphs of *C. jejuni* and *Xcc*, the smooth frequency curves of W (A and T) and S (G and C) nucleotides are distinctly separated as GC% the chromosome are toward both extremes. The distribution was studied by the KS test and the results of the four chromosomes are shown in Fig. 1A(ii, iii), B(ii, iii), C(ii, iii), D(ii, iii) and E(ii, iii). The graphs generated by the KS test suggest the complete overlapping between the complementary nucleotides in the chromosomes except the one of *X. fastidiosa* strain, which is in concordant with the smooth frequency curve. The distributional similarity between the complementary nucleotides is called as ISFDP. A total of 112 bacterial, 49 archaea and 18 eukaryotic chromosomes (Tables 1–3) were analyzed by the KS test to study ISFDP. The *P*-values between the A and T distributions as well as between the G and C distributions are given in Tables 1–3.

Out of 112 bacterial chromosomes, 60 chromosomes exhibited ISFDP, 16 chromosomes exhibited violation between S nucleotides as well as between W nucleotides, 30 chromosomes exhibited violation only between S nucleotides and 7 chromosomes exhibited violation only between W nucleotides (Table 4). Chromosomes of *Alkaliphilus oremlandii* OhILAs (36.26%), *Agrobacterium tumefaciens* C58 (circular; 59.38%), *Mycobacterium ulcerans* Agy99 (65.47%), *Staphylococcus epidermidis* ATCC 12228 (32.1%) and *X. fastidiosa* 9a5c (52.68%) exhibited strong violations between S nucleotides as well as between W nucleotides. Chromosomes of the three *Bacillus anthracis* (35.35%) strains, *Lactobacillus reuteri* F275 (38.87%), *Magnetococcus* sp. MC-1 (54.17%), *Mycobacterium leprae* TN (57.8%), *Rhizobium leguminosarum* bv. *viciae* 3841 (61.09%) and *Rickettsia bellii* RML369-C (31.65%) exhibited strong violation between S nucleotides as well as weak violation between W nucleotides. Chromosomes of *Coxiella burnetii* Dugway 7E9-12 (42.44%) and *Staphylococcus haemolyticus* JCSC1435 (32.79%) exhibited weak violation between S nucleotides as well as between W nucleotides. Chromosome of *Vibrio cholerae* O395 (47.78%) exhibited strong violation of ISFDP only between W nucleotides. Similarly, there are six chromosomes where weak violations only between W nucleotides were observed. Chromosomes of *Bacillus thuringiensis* serovar konkukian 97-27 (34.41%), *Bordetella parapertussis* 12822 (68.1%), *Bordetella pertussis* Tohama 1 (67.72%), *Haemophilus influenzae* PittGG (38.01%), *Helicobacter hepaticus* ATCC 51449 (35.93%), *Lactobacillus acidophilus* NCFM (34.72%), *Lactobacillus brevis* ATCC 367 (46.22%), *Nitrobacter winogradskyi* Nb-255 (62.05%), *Ralstonia solanacearum* GMI1000 chromosome (67.04%), *Rhizobium etli* CFN 42 (61.27%), *Thermotoga maritima* MSB8 (46.25%) and *Thermotoga petrophila* RKU-1 (46.09%) exhibited strong violation only between S nucleotides. Similarly there are 17 chromosomes exhibited weak violation only between S nucleotides. An interesting finding that came from this study is that violations of ISFDP within a chromosome with respect to S and W nucleotides may not be of similar magnitudes. This study suggests that although ISFDP is commonly observed among chromosomes, its violation is not as rare as described earlier.[13] ISFDP violation found in bacteria belongs to different groups, possessing different GC% and with different genome sizes.

Usually, different strains within a species are found to be similar with respect to ISFDP such as the eight *E. coli* strains were observed to exhibit ISFDP between S nucleotides as well as between W nucleotides, the three *B. anthracis* strains are found to be

Table 4. Summary of ISFDP violations in chromosomes of Bacteria, Archaea and Eukaryotes

Organism	Number of chromosomes	Number of chromosomes exhibiting ISFDP for both W and S	Number of chromosomes violating* ISFDP for both W and S	Number of chromosomes violating ISFDP only between S nucleotides	Number of chromosomes violating ISFDP only between W nucleotides
Bacteria	112	60	15 ($5^a+8^b+0^c+2^d$)	30 (13^e+17^f)	07 (1^g+6^h)
Archaea	49	30	06 ($2^a+2^b+2^c+0^d$)	06 (0^e+6^f)	07 (2^g+5^h)
Eukaryotes	18	15	01 ($0^a+0^b+0^c+1^d$)	01 (1^e+0^f)	01 (0^g+1^h)

*Violation of ISFDP includes both weak ($10^{-2} > P \geq 10^{-4}$) and strong ($P < 10^{-4}$).

[a]Strong violation between S nucleotides as well as between W nucleotides.

[b]Strong violation between S nucleotides but weak violation between W nucleotides.

[c]Weak violation between S nucleotides but strong violation between W nucleotides.

[d]Weak violation between S nucleotides as well as between W nucleotides.

[e]Strong violation only between S nucleotides.

[f]Weak violation only between S nucleotides.

[g]Strong violation only between W nucleotides.

[h]Weak violation only between W nucleotides.

similar in terms of their ISFDP violation (strong violation of ISFDP between S nucleotides as well as weak violations of ISFDP between W nucleotides). However, variation among the strains of a bacterial species with respect to ISFDP was observed as follows: out of the two strains of *C. burnetii*, Dugway 7E9-12 strain violated ISFDP, whereas RSA 493 strain exhibited ISFDP. Out of the four *H. influenza* strains, 86-028NP and PittEE exhibited violation of ISFDP, whereas PittGG and Rd KW20 exhibited strong and weak violations only between S nucleotides, respectively. *Xylella fastidiosa* 9a5c exhibited strong violation of ISFDP, whereas *X. fastidiosa* Temecula 1 exhibited weak violation of ISFDP only between S nucleotides. These are called as intra-species ISFDP violations. Chromosomes of four species of *Mycobacterium* genus exhibited a large difference among each other with respect to ISFDP. Chromosome of *Mycobacterium* sp. KMS (68.44%) exhibited parity between S nucleotides as well as between W nucleotides, whereas chromosome of *M. ulcerans* Agy99 (65.47%) exhibited strong violation of the parity between S nucleotides as well as between W nucleotides.

ISFDP in chromosomes of archaea and eukaryotes

Out of the 49 archaea chromosomes, 30 exhibited ISFDP, 6 exhibited violations of it between S nucleotides as well as between W nucleotides, 6 exhibited violations only between S nucleotides and 7 exhibited violations only between W nucleotides (Table 4). Chromosomes of *Methanobrevibacter smithii* ATCC 35061 (31.02%) and *Staphylothermus marinus* F1 (35.71%) exhibited strong violation of ISFDP between S nucleotides as well as between W nucleotides. Chromosomes of *Metallosphaera sedula* DSM 5348 (46.21%) and *Pyrobaculum aerophilum* IM2 (51.34%) exhibited strong violations between S nucleotides but weak violations between W nucleotides. Strong violation between W nucleotides and weak violation between S nucleotides were observed in chromosomes of *Methanosarcina barkeri* Fusaro (39.27%) and *Nitrosopumilus maritimus* SCM1 (31.15%). This suggests that within a chromosome, the magnitude of parity violation between S nucleotides may be different from that between W nucleotides in archaea also like that of bacteria. Intra-species parity violation was also observed in archaea in the case of *Methanococcus maripaludis*. The C5 strain exhibited ISFDP violation between W nucleotides but exhibited parity between S nucleotides. The C6, C7 and S2 strains exhibited ISFDP between S nucleotides as well as between W nucleotides.

Out of the 18 eukaryotic chromosomes belonging to five species, 15 chromosomes exhibited ISFDP (Table 4). Strong violation of ISFDP only between S nucleotides is observed in *Leishmania major* Friedlin chromosome 01 (62.84%). *Plasmodium falciparum* 3D7chromosome 05 exhibited weak violation of parity between S nucleotides as well as between W nucleotides, whereas chromosome 01 exhibited violation of parity only between W nucleotides. The other four chromosomes of *P. falciparum* exhibited parity between S nucleotides as well as between W nucleotides. Similarly, the eight chromosomes of *Saccharomyces cerevisiae* even though exhibited parity between S nucleotides as well as between W nucleotides, the *P*-values either for S nucleotides or for W nucleotides is of more than 10-fold difference among the chromosomes. This differential ISFDP violation observed among chromosomes of an organism suggests that there may not be any strict rule inside a cell to maintain ISFDP.

ISFDP between complementary oligonucleotides in chromosomes

ISP between compositional abundance values of complimentary oligonucleotides is well reported. We studied here the frequency distribution of complementary di- and trinucleotides in chromosomes as described for mononucleotides. The smooth curves of oligonucleotide frequencies have been shown in Supplementary data. In Supplementary Fig. S1a and b, the frequency distributions of dinucleotides have been shown for *E. coli* K12 MG1655 and *Pseudomonas entomophila* L48 chromosome (64.16%). Out of the 12 smooth frequency curves (four palindromic dinucleotides were excluded), overlapping of the curves between complementary dinucleotides is observed. In Fig. 2, though the abundance values of aa, tt, tg and ca dinucleotides in *E. coli* chromosome are close, the distributions between the

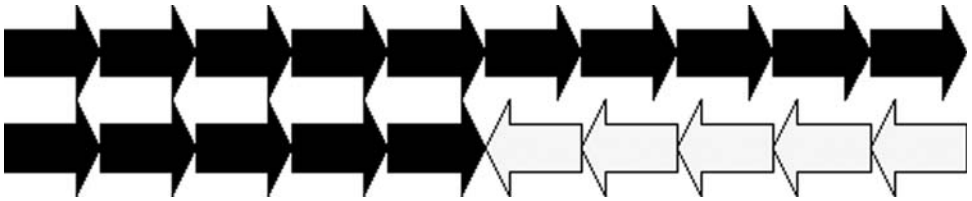

Figure 2. A schematic representation of coding sequence arrangement studied. In the upper row, the entire DNA strand is composed of forward encoded sequences (black color). Parity is not observed in this case. In the lower row, the DNA strand is made up of 50% forward encoded sequences and the other 50% is the reverse encoded sequences (white color). Parity is observed in this case.

complementary dinucleotides are found only overlapping and that of the non-complementary ones are different. The distributions for aa and tt follow a higher standard deviation (values not shown) than that of tg and ca. Similarly, gg and cc dinucleotides distributions exhibit a higher standard deviation (values not shown) than that of the dinucleotides tc and ga, although the abundance values of the four dinucleotides are close to each other. The significance of the similarity was studied by the KS test which suggested that the frequency distributions between complementary dinucleotides are statistically similar. Apart from this, dinucleotides distribution parity has been studied in three more bacterial chromosomes, two archaea chromosomes and one eukaryotic chromosome (data not shown) and similar result has been observed. In Supplementary Fig. S2i and ii, the distribution of 22 trinucleotides of *E. coli* K12 MG1655 chromosome is shown. Like dinucleotides, overlapping between the distributions of complementary trinucleotides is also observed. Distribution similarity between complementary trinucleotides was studied by the KS test for the 64 trinucleotides which suggested that the distributions of complementary trinucleotides within a strand are similar. The same study was done in one more bacterial chromosome (data not shown) and similar results were obtained. Although we did not analyzed the chromosomes of archaea and eukaryotes for trinucleotide distribution parity, it is expected to be there because these chromosomes had exhibited ISFDP for mononucleotides as well as dinucleotides.

ISFDP weakly correlates with Chargaff's second parity

Comparison of ISFDP was done with the ATS/GCS in chromosomes to find out whether one can define the other. GCS was compared with ISFDP violation between S nucleotides and ATS was compared with ISFDP violation between W nucleotides. Among the bacterial chromosomes, maximum GCS was found in *X. fastidiosa* 9a5c with the value 0.0529. All of the 16 chromosomes with GCS $\geq$0.01 were found to violate ISFDP (14 strongly violated and 2 weakly violated). Out of the 18 chromosomes with GCS $\geq$0.005 but $<$0.01, 6 exhibited insignificant violation, 7 exhibited strong violation and 5 exhibited weak violation of ISFDP. Similarly, out of 56 chromosomes with GCS $\geq$0.001 but $<$0.005, 5 exhibited strong violation, 11 exhibited weak violation and 40 exhibited insignificant violation. Out of the 22 chromosomes with GCS $<$0.001, except *B. thuringiensis* Al Hakam chromosome (with GCS value 0.00081 exhibited weak violation of ISFDP) all other exhibited insignificant violation. Maximum ATS was found in *X. fastidiosa* 9a5c with the value 0.04727. Out of the five chromosomes with ATS $\geq$0.01, four were found to violate ISFDP (two strongly violated and two weakly violated), whereas *Mycoplasma hyopneumoniae* J exhibited insignificant violation (with ATS 0.0102). Out of the 14 chromosomes with ATS $\geq$0.005 but $<$0.01, 6 exhibited insignificant violation, 3 exhibited strong violation and 5 exhibited weak violation of ISFDP. Out of the 67 chromosomes with ATS $\geq$0.001 but $<$0.005, 57 exhibited parity, 1 strongly violated and 9 violated weakly between the W nucleotides. All the 26 chromosomes with ATS $\leq$0.001 exhibited insignificant violation of ISFDP. These results suggest that chromosomes with high ATS/GCS ($\geq$0.01) have a stronger propensity to violate ISFDP and chromosomes with low ATS/GCS ($\leq$0.001) have a stronger propensity to exhibit ISFDP. However, chromosomes with intermediate ATS/GCS ($\geq$0.001 and $\leq$0.01) have the possibility of either exhibiting parity or violating the parity.

Correlation analysis was done between the *P*-values (from the KS test between) of W nucleotides and ATS as well as between the *P*-values (from the KS test between) of S nucleotides and GCS. The *r*-values are -0.5572 and -0.4526 for W and S nucleotides, respectively. This suggests that the correlation between the two ISP features is weak. The correlation between ATS and GCS is 0.629, which suggests that parity violation between S nucleotides weakly correlates with parity violation between W nucleotides within a chromosome. Unlike ATS and GCS correlation, no correlation was found between the *P*-values (the KS test) of W nucleotides and that of S nucleotides, which supports that ISFDP and Chargaff's second parity are not the same.

In the case of the archaea chromosomes, the ISFDP analysis revealed similar results to that of bacterial chromosomes. Maximum GCS with the value 0.03768 was found in the chromosome of *M. smithii* ATCC 35061 (31.02%) followed by the value 0.02726 in *S. marinus* F1 (35.71%), in which significant ISFDP violation was also observed. In the GCS interval $0.005 < \text{GCS} \leq 0.01$, there were eight chromosomes out of which five exhibited weak violation and three exhibited insignificant violation of ISFDP. Out of the 24 chromosomes in the interval $0.001 < \text{GCS} \leq 0.005$, 2 exhibited weak violation and 22 exhibited insignificant violation of ISFDP. These results suggest that chromosomes with high ATS/GCS ($\geq$0.01) are most likely going to violate ISFDP and chromosomes with low ATS/GCS ($\leq$0.001) are most likely going to exhibit ISFDP. However, chromosomes with intermediate ATS/GCS (($\geq$0.001 and$\leq$0.01) have the possibility of either exhibiting parity or violating the parity. Pearson's correlation coefficient between ATS and GCS was found to be 0.707847, which is similar to that of the bacterial

analysis. The *r*-values between ATS and the *P*-values of KS (W) as well as GCS and the *P*-values of KS (S) were found to be -0.57495 and -0.47557, respectively, suggesting a weak correlation.

The chromosomes with asymmetric replication topography are more prone to ISFDP violation in bacteria

Bacterial chromosome is a single replicon. Owing to the bidirectional mode of replication, one part of a strand is synthesized as LeS whereas the other part is synthesized as LaS. In most of the chromosomes, the mutational strand asymmetry causes K nucleotides $>$ M nucleotides in LeS and the reverse in (K nucleotides $<$ M nucleotides) in LaS. In an ideal case where the termination site is located symmetrically with respect to the origin of replication in a chromosome, the excess of K nucleotides in LeS will be similar to the excess of M nucleotides in LaS and therefore will cancel each other to exhibit Chargaff's second parity in chromosomes. Potential replication origin and termination sites for different chromosomes based on ATS, GCS, coding sequence skew, nucleotide skew at the third position of codons and oligonucleotides skew in chromosomes have been reported,[31,32] which has been reviewed in detail.[33] Out of the 112 bacterial chromosomes analyzed in this study, information regarding the potential site for the origin and termination of 56 chromosomes is available. ISFDP violation between S nucleotides was compared with the angular deviation of termination site because G $>$ C in LeS is a more universal feature of chromosomes than T $>$ A in LeS. Of the 112 chromosomes, maximum angular deviation of 71.28° is reported in *B. pertussis* Tohama 1. Out of the 14 chromosomes where $\geq$20° angular deviation was observed, 12 exhibited violation of ISFDP between S nucleotides. *Pseudomonas putida* F1 (61.86%) with 36.8° and *C. burnetii* RSA 493 (42.66%) with 31.14° angular deviations exhibited insignificant parity violation. Out of the 11 chromosomes with deviation $\geq$10° but $<$20°, 4 chromosomes exhibited ISFDP violation between S nucleotides. Out of the 30 strains with deviation $\geq$1.0° and $\leq$10°, 9 chromosomes exhibited parity violation between S nucleotides. *Chlamydophila abortus* S263 with angular deviation only 0.569°, parity violation was observed only between S nucleotides. This study indicates that chromosomes with higher asymmetric topography are more prone to violate the parity. However, chromosomes with symmetric replication topography were also observed to violate the parity.

The correlation coefficient between angular deviations and GCS as well as ATS values are 0.474 and 0.357, respectively, suggesting a weak correlation. The correlation between angular deviations and *P*-value of S (the KS test between S nucleotides) as well as that of W (the KS test between W nucleotides) are -0.259 and -0.048, respectively. The angular deviation in *X. fastidiosa* 9a5c is 62.96°, whereas the same in Temecula 1 is 6.44°. The difference in the magnitude of ISFDP violation between the strains might be attributed to the chromosome topography. Comparison for the four *H. influenzae* strains could not be done due to the unavailability of information for all the strains. The Rd KW20 chromosome (that violated ISFDP) has the angular deviation 46°, which might be an important factor to violate ISFDP. Archaea chromosomes have been reported to have more replication origin like eukaryotic chromosomes. Therefore, replication topography will not be applicable to study ISFDP violations in these cases.

Composition of forward- and reverse-encoded sequences within DNA strands might influence the parity

Most of the regions in prokaryotic chromosomes are composed of coding sequences. Presence of both forward- and reverse-encoded sequences in bacterial chromosomes has been proposed for the observation of Chargaff's second parity in chromosomes.[8,9] So we analyzed only coding sequences in chromosomes of bacteria and archaea to study ISFDP as follows (Fig. 2): in one way (Case I), a DNA strand is only composed of only forward-encoded sequences, and in the other way (Case II), a DNA strand is composed of 50% forward-encoded and 50% reverse-encoded sequences. The result is shown for *E. coli* chromosome (Fig. 3A and B). The smooth frequency curves of complementary nucleotides overlap in Fig. 3B, whereas in Fig. 3A, they do not overlap. The significance of these overlaps were studied by the KS test which suggests that the similarity between the distribution of complementary nucleotides in Case II. Similar results were obtained by the analysis of several bacterial (10) and archaea (15) chromosomes.

A comparative analysis between the Ws and Cs in a chromosome with respect to their composition of forward-encoded sequences was done in *X. fastidiosa* species as well as in *H. influenza* species. The relative differences in the compositional abundance values of forward sequences in Ws and Cs of *X. fastidiosa* 9a5c and *X. fastidiosa* Temecula 1 chromosomes are 0.078 and 0.015, respectively, which indicate that the proportion of forward- and reverse-encoded sequence in 9a5c strain is more disproportionate than that of Temecula 1 strain, which might be the reason for a stronger parity violation in the former. The relative differences of the compositional abundance values of forward-encoded sequences in Ws and Cs of *H. influenzae* 86-028NP (exhibits parity) and *H. influenzae* Rd KW20 (violates parity) chromosomes are 0.030 and 0.005, respectively, which suggest that the proportion of forward- and

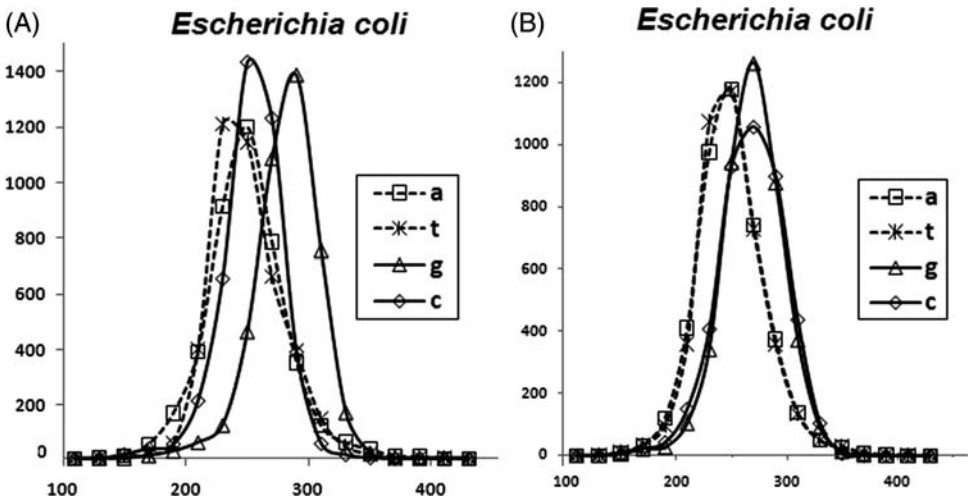

Figure 3. (A and B) Frequency distribution study of nucleotides in coding sequences. Smooth curves present the group-frequency distribution of the four nucleotides a (square), t (asterisk), g (triangle) and c (rhombus). The *X*-axis represents the abundance values of the nucleotide spanning a range, whereas the *Y*-axis represents the frequency of the abundance values. In (A), the frequency of the nucleotides in a DNA strand only composed of forward encoded sequences of *E. coli* is shown (coding sequences analyzed for other chromosomes exhibited the similar feature). It is evident from (B) that the frequency distributions of the complementary nucleotides do not overlap. In (B), the frequency of the nucleotides of the same DNA strand done where 50% of the sequence was joined with the rest after reverse complementation (see the Materials and methods section). This resembled a strand composed of 50% forward encoded sequences and 50% reverse encoded sequences. It is evident from the figures that parity between the complementary nucleotides is observed in this case. These observations have been confirmed by the KS test.

reverse-encoded sequences in 86-028NP strain is more disproportionate than that of Rd KW20 strain. This is in contrast to the result of *X. fastidiosa*, i.e. parity violation is observed in the strain (Rd KW20) with more proportionate gene distribution between Ws and Cs, whereas insignificant parity violation is observed in chromosome with disproportionate gene distribution between the strands. A quantitative estimation of the coding sequences in both the strands of the chromosomes was done in few other bacteria and archaea such as *A. tumefaciens*, *B. subtilis*, *E. coli*, *M. smithii* and *S. marinus* (Fig. 4). Maximum difference of ORF numbers between Ws and Cs was found in *S. marinus*, in which the parity violation was also observed. However, the relative difference of ORFs between the strands is found more in *B. subtilis* than in *M. smithii*. The former exhibited the parity whereas the latter violated it. *Agrobacterium tumefaciens* was shown to possess minimum relative difference of ORF numbers between the strands but violates parity. The results from this indicate that a higher disproportionate composition of forward- and reverse-encoded sequences within a strand has greater propensity to parity violation. However, proportionate composition of the sequences not necessarily implies the exhibition of parity.

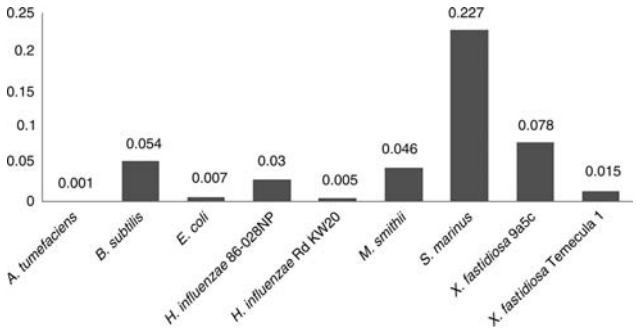

Figure 4. Relative disproportionate composition of ORFs between Ws and Cs in chromosomes. The composition of ORFs in Ws and Cs of seven bacteria and two archaea was studied. Relative disproportionate composition was found out by deducting the ORF numbers between the two strands and then dividing the value obtained by the total number of ORFs present in both the strands. In *A. tumefaciens*, relative disproportionate value found to be minimum suggesting that the difference in the number of ORFs between the strands is relatively minimum when compared with others. In the archaea *S. marinus*, the value is found to be maximum among these nine strains. Both *A. tumefaciens* and *S. marinus* exhibited ISFDP violations, whereas insignificant ISFDP violation observed between *E. coli* and *B. subtilis*. Comparison between the strains of *X. fastidiosa* and *H. influenzae* is shown.

Discussion

We have described in this study a new ISP feature in chromosomes, which is found in bacteria, archaea and eukaryotes. The methodology used to study this parity gives the statistical significance of similarity between the two distributions of complementary nucleotides/oligonucleotides. The basic qualitative feature of ISFDP is not changing for a chromosome even the segmentation is done at randomly taking any point out of the first 1000 nucleotides as the

starting point. In other words, the sampling fluctuation is not affecting the feature. The correlation between the ISFDP and ISP is not strong, which is in accordance with the view that similarity in the total abundance values of two complementary nucleotides will not always yield similarity in their frequency distribution pattern. However, violation of ISP will definitely exhibit violation of ISFDP. Around 50% of the chromosomes in bacteria are found to exhibit ISFDP violations. Chromosomes of *H. influenzae* Rd KW20, *M. tuberculosis* F11, etc., which have been reported to exhibit ISP, are found to violate ISFDP.[27]

ISFDP violation observed in all possible combinations in chromosomes: (i) violation of parity between S nucleotides as well as between W nucleotides; (ii) only between S nucleotides and only between W nucleotides. The correlation between ATS and GCS is found to be not strong suggesting that parity violation between S nucleotides not necessarily always associate with parity violations between W nucleotides and *vice versa*. This can be called as intra-chromosomal parity violations. ISFDP violations of different magnitudes were found among chromosomes of different strains belonging to a species which can be referred as intra-species parity violations. Examples are *C. burnetii*, *H. influenzae* and *X. fastidiosa*. These intra-chromosomal and intra-species violations suggest that there may not be any strict rule existing in cells to maintain ISFDP in chromosomes. Differential ISP among chromosomes within a species and between chromosomes within a bacterium has already been reported in *Chlamydophila pneumoniae* strains and *Deinococcus radiodurans* R1 chromosomes,[27] respectively. However, these were not considered significant in their study due to the lack of statistical proof. Oligonucleotide skew patterns also have been found to be variable among strains of *Yersinia pestis*. These intra-species variations in the chromosomal features are interesting and need in-depth analysis of the genome sequences to find out the reason that might reveal the reason for ISP/ISFDP violation in chromosomes and between the two ISP features.

Enrichment of LeS with K nucleotides over M nucleotides and the *vice versa* in LaS due to the mutational strand asymmetry is a general observation in chromosomes. Owing to the bidirectional replication, GCS/ATS in LeS is cancelled with GCS/ATS in LaS which results in the establishment of parity in chromosomes. The cancellation effect indirectly suggests that the compositional abundance values between the two complementary nucleotides even though they differ within a sub-chromosomal region. This is in support of the observation here that chromosomes with higher GCS/ATS values are violating ISFDP and chromosomes with lower GCS/ATS are exhibiting the parity. However, the chromosomes with intermediate range GCS/ATS are found to exhibit parity as well as violate parity and this violation is independent of genome GC%. For example, *Streptococcus mutans* UA159, *Rickettsia conorii* Malish 7, *C. jejuni* subsp. *jejuni* 81116, *Campylobacter concisus* 13826 and *Lactococcus lactis* subsp. *cremoris* MG1363, *Helicobacter pylori* J99 are (all AT-rich organisms) chromosomes with GCS$\geq$0.005 that exhibit ISFDP between S nucleotides, whereas chromosomes of *B. anthracis* strains (AT rich) with similar GCS ($>$0.005) violate ISFDP between S nucleotides. So ISFDP in these chromosomes is an interesting aspect of future research.

In concordant with the view of the bidirectional replication and establishment of parity in chromosomes, several chromosomes with higher asymmetric replication topography were found to violate ISFDP. The exceptions are *P. putida* F1 and *C. burnetii* RSA 493 chromosomes with 36° and 31° angular deviations, respectively. Chromosomes of *C. abortus* S263 and *Magnetospirillum magneticum* AMB-1, with very less angular deviations 0.57° and 2.14°, respectively, are found violating ISFDP. This indicates that features apart from the replication topography might contribute to the parity establishment in chromosomes. Proportionate composition of forward-encoded sequences between the two strands though thought to be responsible to establish the parity after the analysis of artificially constructed chromosomes, several observations went against it. The extreme case is *A. tumefaciens* where the composition is very much proportionate but violations of ISFDP are strong. So the two factors such as asymmetric replication topography and disproportionate composition of forward-encoded sequences between the strands in chromosomes that were assumed to play important roles in determining ISFDP violations were found to be insufficient.

In spite of different selection/mutation pressures on chromosomes as exemplified by codon usage,[34] replication topography,[31] isochores[35] and GCS/ATS,[21] the tendency of the chromosomes of all types toward maintaining the ISFDP is interesting. Since ISFDP and ISP are the outcomes of compositional abundance of nucleotides (mono/oligo), theories proposed for ISP might hold true for ISFDP. The Nussinov–Forsdyke hypothesis is that stem–loop potential has an adaptive advantage, and therefore an important factor driving the compositional symmetry (ISP) between the complementary oligonucleotides[36,37] has been challenged recently by Chen and Zhao[38] for human chromosomes. This indicates that the stem–loop (recombination) hypothesis might not be the only explanation for ISP in chromosomes. Baisnée *et al*.[8] have argued that the reverse complement symmetry does not result only from point mutation or from

recombination, but from a combination effect of different mechanisms at different orders.[8] Two independent reports have theoretically shown that multiple inversion events in chromosomes can establish ISP.[10,39] Though this hypothesis looks fine theoretically, frequent inversion unable to explain the universal observation of opposite GCS/ATS in LeS and LaS,[40] gene distribution asymmetry between the strands[41] and the maintenance of gene orders among different bacterial chromosomes.[42] This hypothesis also does not describe any functional significance/advantage of the ISP/ISFDP feature, which is so wide spread in chromosomes. Theoretically, it has also been argued that the mismatch error repairing system is responsible to establish Chargaff's second parity rule in chromosomes.[13] However, the intra-chromosomal parity violation observed in eukaryotes (this study) goes against this hypothesis.

We think the important factor that determines ISP/ISFDP in chromosomes is the bidirectional replication. This causes one part of a strand Ws/Cs as LeS and the other part as LaS. The strand mutational asymmetry and gene distribution asymmetry between LeS and LaS therefore cancel out each other within the strand to exhibit the parity. In the case of ssDNA/ssRNA viruses, gene distribution is restricted to one strand only depending on which these are called as either plus or − strand viruses. The genome size is also not large (<10 kb) in these phages[43,44] and during replication, one strand only acts as the template on which the other strand is made. Most likely these features are responsible for violating the parity in these genomes. The advantages of bidirectional replication in bacteria and archaea where the nucleus is absent are as follows: (i) quicker completion of replication than the unidirectional mode of replication and (ii) the meeting of the two replication forks might be sending some signal to the cell for the completion of chromosome replication where the nucleus is absent. Symmetric replication topography will help to terminate the replication from the origin in a lesser time in comparison with an asymmetric topography. So the selection pressure to maintain the symmetric replication topography in fast-growing bacteria is likely to be more than that in slow-growing bacteria. This proposition has similarity with the Selection Mutation Drift theory proposed for codon usage[45] in bacteria. Our study of ISFDP of *Vibrio* species (the generation time is 0.2−0.3 h; fast-growing) in this context seems to be also not holding true here because its chromosomes violate ISFDP between W nucleotides. Moreover, comparison of generation time[40] with asymmetry in replication topography of chromosomes[32] exhibits no correlation (data not shown). More research on this aspect will give a conclusive result if the growth rate has any relation with parity establishment in chromosomes. In conclusion, our study has revealed an interesting aspect of ISP. Future research will reveal the reason for the presence of this parity in chromosomes.

Acknowledgements: A part of this work has been presented as a poster by S.K.R. in the International Conference ISMB2008 at Toronoto, Canada. Support to S.K.R. from DST (India), INSA (India) and Tezpur University for attending this conference is thankfully acknowledged. We thank the department of Biotechnology, Govt. of India for awarding MSc student fellowships to A.K. and P.K.J. The authors thank to J. R. Lobry for his critical comments on the work and are very much grateful to D. R. Forsdyke (the reviewer of the manuscript) for his comments on the manuscript.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

References

1. Chargaff, E. 1951, Structure and function of nucleic acids as cell constituents, *Fed. Proc.*, **10**, 654−9.
2. Forsdyke, D.R. and Mortimer, J.R. 2000, Chargaff's legacy, *Gene*, **261**, 127−37.
3. Watson, J.D. and Crick, F.H.C. 1953, Molecular structure of nucleic acids: a structure for deoxyribonucleic acid, *Nature*, **171**, 737−8.
4. Karkas, J.D., Rudner, R. and Chargaff, E. 1968, Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerases, *Proc. Natl Acad. Sci. USA*, **60**, 915−20.
5. Rudner, R., Karkas, J.D. and Chargaff, E. 1969, Separation of microbial deoxyribonucleic acids into complementary strands, *Proc. Natl Acad. Sci. USA*, **63**, 152−9.
6. Prabhu, V.V. 1993, Symmetry observed in long nucleotide sequences, *Nucleic Acid Res.*, **21**, 2797−800.
7. Qi, D. and Cuticchia, A.J. 2001, Compositional symmetries in complete genomes, *Bioinformatics*, **17**, 557−9.
8. Baisnée, P.F., Hampson, S. and Baldi, P. 2002, Why are complementary DNA strands symmetric?, *Bioinformatics*, **18**, 1021−33.
9. Verma, S.K., Das, D., Satapathy, S.S., Buragohain, A.K. and Ray, S.K. 2005, Compositional Symmetry of DNA duplex in bacterial genomes, *Curr. Sci.*, **89**, 374−84.
10. 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**, 17828−33.
11. Mitchell, D. and Bridge, R. 2006, A test of Chargaff's second rule, *Biochem. Biophys. Res. Commun.*, **340**, 1−5.
12. Nikolaou, C. and Almirantis, Y. 2006, Deviation from Chargaff's second parity rule in organellar DNA insights into the evolution of organellar genomes, *Gene*, **381**, 34−41.

13. Deng, B. 2007, Mismatch repair error implies Chargaff's second parity rule, arXiv:0704.2191v2 [q-bio.GN], 1–9, [http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.2191v2.pdf].
14. Sueoka, N. 1995, Intra-strand parity rules of DNA base composition and usage biases of synonymous codons, *J. Mol. Evol.*, **40**, 318–25.
15. Lobry, J.R. 1995, Properties of a general model of DNA evolution under no-strand-bias conditions, *J. Mol. Evol.*, **40**, 326–30.
16. Frank, A.C. and Lobry, J.R. 1999, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, *Gene*, **238**, 65–77.
17. Nikolaou, C. and Almirantis, Y. 2005, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, *Nucleic Acid Res.*, **33**, 6816–22.
18. Sueoka, N. 1999, Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position, *Gene*, **238**, 53–8.
19. Rocha, E.P.C., Danchin, A. and Viari, A. 1999, Universal replication biases in bacteria, *Mol. Microbiol.*, **32**, 11–6.
20. Lobry, J.R. and Sueoka, N. 2002, Asymmetric directional mutation pressures in bacteria, *Genome Biol.*, **3**, research0058.1–0058.14.
21. Grigoriev, A. 1998, Analyzing genomes with cumulative skew diagrams, *Nucleic Acid Res.*, **26**, 2286–90.
22. Francino, M.P. and Ochman, H. 1997, Strand asymmetries in DNA evolution, *Trends Genet.*, **13**, 240–5.
23. Johnson, A. and O'Donnell, M. 2005, Cellular DNA replicases: components and dynamics at the replication fork, *Annu. Rev. Biochem.*, **74**, 283–314.
24. Bell, S.J. and Forsdyke, D.R. 1999, Accounting units in DNA, *J. Theor. Biol.*, **197**, 51–61.
25. Francino, M., Chao, L., Riley, M.A. and Ochman, H. 1996, Asymmetries generated by transcription-coupled repair in enterobacterial genes, *Science*, **272**, 107–9.
26. McLean, M.J., Wolfe, K.H. and Devine, K.M. 1998, Base composition skews, replication, and orientation in 12 prokaryote genomes, *J. Mol. Evol.*, **47**, 691–6.
27. Shioiri, C. and Takahata, N. 2001, Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry, *J. Mol. Evol.*, **53**, 364–76.
28. Nikiforov, A.M. 1994, Algorithm AS 288: Exact two-sample Smirnov test for arbitrary distributions, *Appl. Stat.*, **43**, 265–84.
29. Kolmogorov, A. 1941, Confidence limits for an unknown distribution function, *Ann. Math. Stat.*, **12**, 461–3.
30. Smirnov, N.V. 1939, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bull. Moscow Univ.*, **2**, 3–14.
31. Frank, A.C. and Lobry, J.R. 2000, Oriloc: prediction of replication boundaries in annotated bacterial chromosomes, *Bioinformatics*, **16**, 560–1.
32. Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H-H. and Ussery, D.W. 2006, Origin of replication in circular prokaryotic chromosomes, *Environ. Microbiol.*, **8**, 353–61.
33. Sernova, N.V. and Gelfand, M.S. 2008, Identification of replication origins in prokaryotic genomes, *Brief. Bioinform.*, **9**, 376–91.
34. Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. 2005, Variation in the strength of selected codon usage bias among bacteria, *Nucleic Acid Res.*, **33**, 1141–53.
35. Duret, L., Eyre-Walker, A. and Galtier, N. 2006, A new perspective on isochore evolution, *Gene*, **385**, 71–4.
36. Nussinov, R. 1984, Strong doublet preferences in nucleotide sequences and DNA geometry, *J. Mol. Evol.*, **20**, 111–9.
37. Forsdyke, D.R. 1995, A stem–loop 'kissing' model for the initiation of recombination and the origin of introns, *Mol. Biol. Evol.*, **12**, 949–58.
38. Chen, L. and Zhao, H. 2005, Negative correlation between compositional symmetries and local recombination rates, *Bioinformatics*, **21**, 3951–8.
39. Okamura, K., Wei, J. and Scherer, S.W. 2007, Evolutionary implications of inversions that have caused intra-strand parity in DNA, *BMC Genomics*, **8**, 160.
40. Rocha, E.P.C. 2004, The replication-related organization of bacterial genomes, *Microbiology*, **150**, 1609–27.
41. Rocha, E.P.C. and Danchin, A. 2003, Gene essentiality determines chromosome organization in bacteria, *Nucleic Acid Res.*, **31**, 6570–7.
42. Rocha, E.P.C. 2008, The organization of the bacterial genome, *Annu. Rev. Genet.*, **42**, 211–33.
43. Adams, M.J. and Antoniw, J.F. 2005, DPVweb: An open access internet resource on plant viruses and virus diseases, *Outlooks on Pest Management*, **16**, 268–70.
44. Adams, M.J. and Antoniw, J.F. 2006, DPVweb: a comprehensive database of plant and fungal virus genes and genomes, *Nucleic Acids Res.*, **34**, D382–5. http://www.dpvweb.net/.
45. Bulmer, M. 1991, The selection-mutation-drift theory of synonymous codon usage, *Genetics*, **192**, 897–907.