DISTRIBUTION OF DINUCLEOTIDE AND ODDS-RATIO SKEWS

The distributions of dinucleotides are very similar along reverse complementary strands on a chromosomal scale [1] and the absolute values of the corresponding skews, when present, are thought to be rather small and thus insignificant [2]. Here, we assess strand-asymmetries of dinucleotide distributions, at the level of both observed frequencies and odds-ratios, along leading strand and CDS concatenate. Besides dinucleotide frequency skews and odds-ratio skews, we also compute mononucleotide skews as a benchmark for our study. Since Firmicutes are known to have atypical mononucleotide skews, in particular $S_{leading}^{A-T}$ [3], we study the species of this phylum separately. Thus, we divide our dataset into two groups, the one corresponding to Firmicutes (Firmicutes group) and the other to all the remaining bacteria of our collection (Firmicutes-excluded group). Then, for each skew, we examine the distribution of its absolute and signed values along these two groups, and estimate the corresponding quantiles (Table 1, Table 2). Signed skew values show the direction of strand asymmetries, while their absolute values are more indicative of the magnitude of such asymmetries.

A. leading strand

		Firmicutes excluded							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	3.6e-05	0.0023	0.0059	0.01	0.014	0.018	0.029	0.11
	S^{G-C}	0.00025	0.01	0.019	0.03	0.035	0.041	0.062	0.22
	S^{AG-CT}	0.00028	0.0039	0.0079	0.015	0.019	0.027	0.056	0.22
	S^{GA-TC}	0.00014	0.0057	0.011	0.022	0.03	0.039	0.063	0.27
	S^{GG-CC}	0.00018	0.024	0.041	0.061	0.071	0.084	0.11	0.33
	S^{AA-TT}	3.9e-05	0.0038	0.0072	0.016	0.023	0.029	0.06	0.21
	S^{AC-GT}	0.00018	0.014	0.029	0.046	0.055	0.066	0.1	0.34
	S^{CA-TG}	0.00015	0.017	0.03	0.044	0.052	0.062	0.092	0.3
dinucleotide odds-ratio skews	P^{AG-CT}	0.00029	0.0021	0.0042	0.012	0.016	0.019	0.028	0.13
	P^{GA-TC}	6.3e-05	0.0031	0.0059	0.014	0.017	0.02	0.03	0.1
	P^{GG-CC}	0.00014	0.0022	0.0052	0.0094	0.013	0.019	0.042	0.26
	P^{AA-TT}	0.00049	0.0047	0.0084	0.014	0.02	0.025	0.036	0.09
	P^{AC-GT}	1.8e-05	0.0024	0.0045	0.011	0.017	0.022	0.034	0.13
	P^{CA-TG}	0.00013	0.0027	0.005	0.012	0.016	0.021	0.034	0.12

		Firmicutes							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	0.0012	0.012	0.018	0.03	0.037	0.047	0.055	0.11
	S^{G-C}	0.04	0.077	0.086	0.095	0.099	0.11	0.13	0.22
	S^{AG-CT}	0.0034	0.078	0.096	0.11	0.13	0.15	0.22	0.3
	S^{GA-TC}	0.063	0.12	0.13	0.14	0.15	0.19	0.21	0.36
	S^{GG-CC}	0.058	0.15	0.15	0.18	0.2	0.21	0.24	0.36
	S^{AA-TT}	0.0048	0.017	0.025	0.04	0.06	0.068	0.087	0.19
	S^{AC-GT}	0.002	0.009	0.039	0.057	0.059	0.061	0.092	0.12
	S^{CA-TG}	0.034	0.047	0.051	0.066	0.068	0.074	0.096	0.19
dinucleotide odds-ratio skews	P^{AG-CT}	0.0036	0.0064	0.0089	0.031	0.04	0.057	0.075	0.1
	P^{GA-TC}	0.0017	0.02	0.034	0.049	0.055	0.057	0.074	0.17
	P^{GG-CC}	8e-06	0.0013	0.0081	0.025	0.034	0.04	0.051	0.24
	P^{AA-TT}	0.00063	0.018	0.026	0.036	0.041	0.043	0.062	0.14
	P^{AC-GT}	2e-04	0.0073	0.012	0.015	0.017	0.019	0.037	0.11
	P^{CA-TG}	0.00048	0.003	0.01	0.033	0.035	0.036	0.038	0.13

B. CDS concatenate

		Firmicutes excluded							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	0.00014	0.004	0.0073	0.016	0.023	0.028	0.048	0.12
	S^{G-C}	5.1e-05	0.0065	0.011	0.023	0.037	0.05	0.078	0.19
	S^{AG-CT}	1.4e-05	0.021	0.044	0.074	0.09	0.1	0.14	0.25
	S^{GA-TC}	0.00013	0.0083	0.016	0.055	0.078	0.093	0.14	0.34
	S^{GG-CC}	0.00039	0.018	0.032	0.064	0.086	0.11	0.16	0.4
	S^{AA-TT}	4.7e-06	0.014	0.031	0.057	0.066	0.077	0.12	0.2
	S^{AC-GT}	0.00059	0.0071	0.019	0.036	0.044	0.051	0.078	0.21
	S^{CA-TG}	0.00071	0.017	0.038	0.058	0.068	0.074	0.11	0.21
dinucleotide odds-ratio skews	P^{AG-CT}	0.0027	0.034	0.064	0.11	0.13	0.15	0.2	0.34
	P^{GA-TC}	4e-05	0.011	0.02	0.048	0.06	0.068	0.091	0.19
	P^{GG-CC}	0.00015	0.0078	0.017	0.031	0.042	0.05	0.075	0.24
	P^{AA-TT}	0.00017	0.012	0.023	0.054	0.077	0.1	0.16	0.29
	P^{AC-GT}	7e-04	0.018	0.029	0.051	0.062	0.076	0.11	0.26
	P^{CA-TG}	0.0046	0.023	0.036	0.065	0.084	0.098	0.16	0.27
		Firmicutes							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	0.00041	0.021	0.029	0.043	0.072	0.078	0.085	0.15
	S^{G-C}	0.016	0.071	0.072	0.089	0.1	0.12	0.15	0.28
	S^{AG-CT}	0.0022	0.038	0.061	0.096	0.13	0.19	0.26	0.33
	S^{GA-TC}	0.071	0.11	0.14	0.15	0.21	0.24	0.26	0.45
	S^{GG-CC}	0.029	0.14	0.15	0.2	0.23	0.27	0.3	0.46
	S^{AA-TT}	0.00099	0.04	0.048	0.063	0.11	0.12	0.13	0.26
	S^{AC-GT}	0.00016	0.0018	0.0031	0.019	0.029	0.039	0.054	0.2
	S^{CA-TG}	0.0064	0.015	0.02	0.066	0.067	0.069	0.11	0.2
dinucleotide odds-ratio skews	P^{AG-CT}	0.00079	0.026	0.045	0.078	0.083	0.089	0.12	0.25
	P^{GA-TC}	0.0093	0.028	0.035	0.058	0.061	0.077	0.09	0.21
	P^{GG-CC}	0.00089	0.0065	0.01	0.058	0.07	0.08	0.088	0.27
	P^{AA-TT}	0.0034	0.024	0.033	0.053	0.073	0.077	0.088	0.17
	P^{AC-GT}	0.032	0.051	0.053	0.06	0.07	0.076	0.09	0.22
	P^{CA-TG}	0.005	0.015	0.031	0.053	0.056	0.065	0.085	0.19

Table 1. Quantiles of skew absolute values, corresponding to the given probabilities. The smallest observation corresponds to a probability of 0% and the largest to a probability of 100%. Skew absolute values measure the magnitude of strand asymmetries (the deviations from intra-strand compositional parities). Our collection is divided into two groups, one corresponding to all examined phyla except Firmicutes (Firmicutes excluded), and the other corresponding only to Firmicutes. Calculations are made along A: the leading strand, and B: the CDS concatenate.

As regards to the leading strand, in the Firmicutes-excluded group the median absolute values (50th quantiles) of dinucleotide skews range from 1.9% (for $|S_{leading}^{AG-CT}|$) up to 7.1% (for $|S_{leading}^{GG-CC}|$), while the median $|S_{leading}^{A-T}|$ is 1.4% and the median $|S_{leading}^{G-C}|$ is 3.5% (Table 1A). The corresponding values are even higher, when only Firmicutes are examined. For instance, the median $|S_{leading}^{GG-CC}|$ in leading strand equals 20%, while the median $|S_{leading}^{G-C}|$ in leading strand is 9.9%. Likewise, the magnitude of odds-ratio skews clearly indicates the presence of strand-asymmetries at the level of nearest-neighbour correlations. For instance, the median $|P_{leading}^{GA-TC}|$ is 1.7% in Firmicutes-excluded group and 5.5% in Firmicutes (Table 1A).

In CDS concatenates, the compositional asymmetries are, in most cases, even more pronounced compared to leading strand (Table 1B). In Firmicutes-excluded group, the median $|S_{CDS}^{AG-CT}|$ in CDS concatenates is 4.7 times the one in leading strand (median $|S_{CDS}^{AG-CT}|$=9%, median $|S_{leading}^{AG-CT}|$=1.9%). The median odds-ratio skew of the same dinucleotide pair, AG-CT, is 6.8 times higher in CDS concatenates (median $|P_{CDS}^{AG-CT}|$=13%) than in leading strand (median $|P_{leading}^{AG-CT}|$=1.6%). Note that, when Firmicutes are excluded, all pairs of reverse complementary dinucleotides, except GA-TC and GG-CC, have median absolute values of odds-ratio skews greater than the ones of their observed frequency skews. In Firmicutes, strong asymmetries are also observed in CDS concatenates. However, the differences of the skews magnitude between CDS concatenate and leading strand are not so strong, given that skews in leading strand are already high in this phylum.

A. leading strand

		Firmicutes excluded							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	-0.11	-0.047	-0.027	-0.015	-0.012	-0.0086	-0.0014	0.098
	S^{G-C}	-0.031	0.0091	0.019	0.03	0.035	0.041	0.062	0.22
	S^{AG-CT}	-0.036	-0.005	0.0021	0.014	0.018	0.027	0.056	0.22
	S^{GA-TC}	-0.038	9e-04	0.0099	0.021	0.03	0.039	0.063	0.27
	S^{GG-CC}	-0.066	0.018	0.04	0.061	0.071	0.084	0.11	0.33
	S^{AA-TT}	-0.21	-0.08	-0.05	-0.026	-0.016	-0.011	0.0024	0.16
	S^{AC-GT}	-0.34	-0.14	-0.1	-0.066	-0.055	-0.046	-0.028	0.038
	S^{CA-TG}	-0.3	-0.12	-0.092	-0.062	-0.052	-0.044	-0.03	0.043
dinucleotide odds-ratio skews	P^{AG-CT}	-0.13	-0.025	-0.019	-0.0042	-0.00073	0.0041	0.02	0.12
	P^{GA-TC}	-0.029	-0.006	0.00079	0.011	0.016	0.02	0.03	0.1
	P^{GG-CC}	-0.26	-0.065	-0.036	-0.011	-0.0047	0.00017	0.0085	0.072
	P^{AA-TT}	-0.09	-0.02	-0.0075	0.0089	0.013	0.019	0.031	0.085
	P^{AC-GT}	-0.13	-0.047	-0.033	-0.019	-0.012	-0.0075	6e-04	0.088
	P^{CA-TG}	-0.084	-0.04	-0.027	-0.012	-0.0079	-0.003	0.011	0.12

		Firmicutes							
		0%	**10%**	**20%**	**40%**	**50%**	**60%**	**80%**	**100%**
mono- and di-nucleotide skews	S^{A-T}	-0.0075	0.012	0.018	0.03	0.037	0.047	0.055	0.11
	S^{G-C}	0.04	0.077	0.086	0.095	0.099	0.11	0.13	0.22
	S^{AG-CT}	0.0034	0.078	0.096	0.11	0.13	0.15	0.22	0.3
	S^{GA-TC}	0.063	0.12	0.13	0.14	0.15	0.19	0.21	0.36
	S^{GG-CC}	0.058	0.15	0.15	0.18	0.2	0.21	0.24	0.36
	S^{AA-TT}	-0.017	0.016	0.025	0.04	0.06	0.068	0.087	0.19
	S^{AC-GT}	-0.12	-0.12	-0.092	-0.061	-0.059	-0.057	-0.039	0.059
	S^{CA-TG}	-0.19	-0.11	-0.096	-0.074	-0.068	-0.066	-0.051	-0.034
dinucleotide odds-ratio skews	P^{AG-CT}	-0.1	-0.06	-0.041	-0.0077	-0.0042	0.0096	0.074	0.092
	P^{GA-TC}	-0.0019	0.02	0.034	0.049	0.055	0.057	0.074	0.17
	P^{GG-CC}	-0.24	-0.064	-0.051	-0.038	-0.027	-0.016	0.00086	0.04
	P^{AA-TT}	-0.14	-0.072	-0.062	-0.043	-0.041	-0.036	-0.025	0.031
	P^{AC-GT}	-0.015	-0.0091	0.0066	0.015	0.017	0.019	0.037	0.11
	P^{CA-TG}	-0.13	-0.041	-0.036	-0.0047	-0.00052	0.018	0.037	0.08

		B. CDS concatenate							
		Firmicutes excluded							
		0%	10%	20%	40%	50%	60%	80%	100%
mono- and di-nucleotide skews	S^{A-T}	-0.063	-0.025	-0.0083	0.006	0.011	0.022	0.047	0.12
	S^{G-C}	-0.059	-0.012	0.0016	0.017	0.036	0.049	0.078	0.19
	S^{AG-CT}	-0.21	-0.13	-0.11	-0.07	-0.038	-0.0026	0.091	0.25
	S^{GA-TC}	-0.052	-0.0098	0.0066	0.055	0.078	0.093	0.14	0.34
	S^{GG-CC}	-0.15	-0.0069	0.019	0.063	0.085	0.11	0.16	0.4
	S^{AA-TT}	-0.13	-0.019	0.0022	0.054	0.064	0.074	0.12	0.2
	S^{AC-GT}	-0.21	-0.059	-0.037	-0.0026	0.015	0.033	0.06	0.2
	S^{CA-TG}	-0.21	-0.12	-0.1	-0.074	-0.067	-0.058	-0.036	0.12
dinucleotide odds-ratio skews	P^{AG-CT}	-0.34	-0.23	-0.2	-0.15	-0.13	-0.11	-0.064	0.044
	P^{GA-TC}	-0.12	-0.043	-0.0053	0.033	0.054	0.064	0.088	0.19
	P^{GG-CC}	-0.14	-0.047	-0.022	0.012	0.026	0.039	0.066	0.24
	P^{AA-TT}	-0.18	-0.068	-0.027	0.022	0.045	0.087	0.16	0.29
	P^{AC-GT}	-0.035	0.005	0.027	0.051	0.062	0.076	0.11	0.26
	P^{CA-TG}	-0.27	-0.2	-0.16	-0.098	-0.084	-0.062	-0.033	0.067
		Firmicutes							
		0%	10%	20%	40%	50%	60%	80%	100%
mono- and di-nucleotide skews	S^{A-T}	-0.0037	0.021	0.029	0.043	0.072	0.078	0.085	0.15
	S^{G-C}	0.016	0.071	0.072	0.089	0.1	0.12	0.15	0.28
	S^{AG-CT}	-0.092	0.035	0.059	0.096	0.13	0.19	0.26	0.33
	S^{GA-TC}	0.071	0.11	0.14	0.15	0.21	0.24	0.26	0.45
	S^{GG-CC}	0.029	0.14	0.15	0.2	0.23	0.27	0.3	0.46
	S^{AA-TT}	-0.0046	0.04	0.048	0.063	0.11	0.12	0.13	0.26
	S^{AC-GT}	-0.1	-0.055	-0.051	-0.021	-0.016	-0.0017	0.0069	0.2
	S^{CA-TG}	-0.2	-0.12	-0.11	-0.069	-0.067	-0.066	-0.02	0.01
dinucleotide odds-ratio skews	P^{AG-CT}	-0.25	-0.17	-0.12	-0.082	-0.076	-0.037	0.045	0.096
	P^{GA-TC}	-0.028	0.023	0.035	0.058	0.061	0.077	0.09	0.21
	P^{GG-CC}	-0.27	-0.017	0.00022	0.024	0.06	0.072	0.086	0.12
	P^{AA-TT}	-0.17	-0.09	-0.088	-0.077	-0.069	-0.037	-0.03	0.1
	P^{AC-GT}	0.032	0.051	0.053	0.06	0.07	0.076	0.09	0.22
	P^{CA-TG}	-0.19	-0.11	-0.085	-0.054	-0.044	-0.0085	0.037	0.072

Table 2. Quantiles of skews, corresponding to the given probabilities. The smallest observation corresponds to a probability of 0% and the largest to a probability of 100%. Negative skew values are highlighted with red. Our collection is divided into two groups, one corresponding to all examined phyla except Firmicutes (Firmicutes excluded), and the other corresponding only to Firmicutes. Calculations are made along A: the leading strand, and B: the CDS concatenate.

Subsequently, we examine the direction of strand asymmetries (Table 2). To highlight our findings, we focus on the AC-GT pair. 80% of the Firmicutes-excluded group has $S_{leading}^{AC-GT}$ that falls between -34% and -2.8% (Table 2A), when less than half of this group takes negative S_{CDS}^{AC-GT} values in CDS concatenates (Table 2B). Thus, the skews can be biased towards different dinucleotides, depending on the strand we examine. Moreover, in Firmicutes, more than 80% have negative $S_{leading}^{AC-GT}$ along the leading strand (the 80^{th} quantile of $S_{leading}^{AC-GT}$ equals -3.9%), while at least 80% of them have positive $P_{leading}^{AC-GT}$ (the 20^{th} quantile of $P_{leading}^{AC-GT}$ equals 0.66%). Thus, for a given pair of reverse complementary dinucleotides, strand asymmetries of its observed frequencies can be biased towards the opposite direction of its odds-ratio skews.

To clarify our findings, let us comment on certain aspects regarding doublet skews in terms of observed frequencies and odds ratios. For example, $S_{leading}^{AC-GT}$ = -3.5% means that along the leading strand there are 3.5% more GT doublets than AC doublets. On the other hand, $P_{leading}^{AC-GT}$ = 3.5% states that the relative abundance of AC is greater than the relative abundance of GT by 3.5% along the leading strand of genes, or, which is the same, AC (GT) odds ratio is higher (lower) in leading than in lagging strand by 3.5%. In other words, given the single-nucleotide composition of DNA, adenine (guanine) and cytosine (thymine) residues show a greater (lower) tendency to form AC (GT) doublets in leading than in lagging strand. However, the base composition of the leading strand is complementary to and generally different from that of lagging strand. Therefore, though the portion of As and Cs that are grouped together in AC doublets is greater in leading than lagging strand, this does not entail a higher frequency of occurrence of ACs lying on the leading versus lagging strand. If the leading strand is more enriched in Gs and Ts than the lagging strand, then $S_{leading}^{AC-GT} < 0$ and $P_{leading}^{AC-GT} > 0$ may simultaneously apply. In fact this is the case for several chromosomes in our collection.

REFERENCES

1. Baisnée, P.-F., Hampson, S., and Baldi, P. 2002, Why are complementary DNA strands symmetric? *Bioinformatics*, **18**, 1021–33.

2. Shioiri, C., and Takahata, N. 2001, Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.*, **53**, 364–76.

3. Charneski, C. A., Honti, F., Bryant, J. M., Hurst, L. D., and Feil, E. J. 2011, Atypical AT skew in Firmicute genomes results from selection and not from mutation. *PLoS Genet.*, **7**, e1002283.