BMC Genomics

BioMed Central

Research article

Pervasive properties of the genomic signature

Robert W Jernigan[1] and Robert H Baran*[2]

Address: [1]Department of Mathematics and Statistics, The American University, Washington, DC 20016 USA and [2]Office of Naval Research, 800 North Quincy Street, Arlington, VA 22217 USA

E-mail: Robert W Jernigan - jernigan@american.edu; Robert H Baran* - baranr@onr.navy.mil

*Corresponding author

Published: 9 August 2002

BMC Genomics 2002, **3**:23

Received: 22 April 2002

Accepted: 9 August 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/23

Abstract

Background: The dinucleotide relative abundance profile can be regarded as a genomic signature because, despite diversity between species, it varies little between 50 kilobase or longer windows on a given genome. Both the causes and the functional significance of this phenomenon could be illuminated by determining if it persists on smaller scales. The profile is computed from the base step "odds ratios" that compare dinucleotide frequencies to those expected under the assumption of stochastic equilibrium (thorough shuffling). Analysis is carried out on 22 sequences, representing 19 species and comprised of about 53 million bases all together, to assess stability of the signature in windows ranging in size from 50 kilobases down to 125 bases.

Results: Dinucleotide relative abundance distance from the global signature is computed locally for all non-overlapping windows on each sequence. These distances are log-normally distributed with nearly constant variance and with means that tend to zero slower than reciprocal square root of window size. The mean distance within genomes is larger for protist, plant, and human chromosomes, and smaller for archaea, bacteria, and yeast, for any window size.

Conclusions: The imprint of the global signature is locally pervasive on all scales considered in the sequences (either genomes or chromosomes) that were scanned.

Background

Compositional heterogeneity pervades the genome on all scales [1] and attempts to partition DNA sequences into homogeneous segments give different results depending on the method chosen, even for genomes as small as the 49 kilobases (kb) of bacteriophage lambda [2]. It therefore seems remarkable that dinucleotide relative abundance profiles exhibit local stability in the sense that, when computed for any 50 kb window on a given microbial genome, the profile is about the same as if computed globally from the bulk genomic DNA of the organism [3,4]. The dinucleotide relative abundance profile, previously called the "general design [5]," can be viewed as a "genomic signature" that reflects a "total net response to selective pressure [6]." Yet the 50 kb window spans the complete genome of bacteriophage lambda and hence it would not be surprising to discover local instability in the profile seen through smaller windows.

Bacterial genomes commonly exhibit an approximate balance between purine (A+G) and pyrimidine (C+T) fractions, and between amino (A+C) and keto (G+T) fractions, when the whole leading strand is examined. Locally, however, these fractions fluctuate from their global averages almost as dramatically as the strong (C+G) and weak (A+T) fractions that do not show global balance. In-

deed, these compositional biases, extending over hundreds of kilobases, are strongly correlated with the direction of replication in some species [7]. Perhaps the stability of the dinucleotide relative abundance profile in 50 kb and longer windows conceals highly variable behavior on a shorter scale. By analogy, the variability of the profile within 50 kb windows could have as much functional significance as invariance between such windows.

The stability of the dinucleotide relative abundance profile could result from constraints on dinucleotide stacking energy and DNA helicity, context-dependent mutation pressures, and replication and repair mechanisms [3,4,6,8]. To gauge the influence of these or other factors in contributing to total net response, it again seems pertinent to ask whether stability persists on smaller scales than 50 kb. If it is reasonable to suppose that signature variation between species is due to differences in replication machinery, operating through local DNA structures and base step conformational tendencies [9], then we may expect to find stability on the scale of the machinery. For example, the proximity of dinucleotide relative abundance profiles in two samples could reflect the similarity of enzymes [10] that engage them in their replication processes. When these samples come from the same genome, which encodes the same enzymes, a close proximity should be expected.

Replicational constraints that act on all parts of the genome could also manifest themselves in codon usage preference. The genomic signature may be associated with codon usage to the extent that it embodies these constraints. Given that the signature "pervades both coding and noncoding DNA," its uniformity throughout the genome "cannot be explained by preferential codon usage [9]." The converse proposition, that codon usage is explained by dinucleotide relative abundance, is hardly more defensible, since dinucleotide frequencies do not determine trinucleotide frequencies [11]. Yet the replicational constraints could be a common factor underlying both phenomena. If the global signature pervades an open reading frame (ORF), the trinucleotide frequencies will to some extent express the signature as modulated by the local composition. (Codon usage may exhibit more diversity within genomes by virtue of its sensitivity to composition.) This kind of association would be supported by discovery of signature stability in windows smaller than the average ORF and negated by the breakdown of signature stability below some window size that far exceeds the typical ORF length, unless protein coding sequences are less sensitive to local replicational constraints than dinucleotide relative abundance [12].

Genomes are identifiable by their signatures and dissimilarity between signatures is used to estimate the evolutionary distance between species [4,6]. If the imprint of the global signature is locally pervasive, down to the scale of the single gene or coding sequence, large deviations on that scale could highlight segments introduced by recent horizontal transfer from another species [13]. So-called filtration methods, based on dissimilarity measures computed from dinucleotide counts, have been employed for the alignment-free computation of evolutionary distances between homologous sequences [14,15]. The "transition matrix method" was a similar technique involving raw counts of amino acid pairs in protein primary sequences [16]. Phylogeny construction based on dinucleotide relative abundance distance ("delta-distance") would seem to have a similar intent although its application has been to whole genomes [9]. Mathematically, however, the use of relative abundance instead of raw abundance (frequencies or counts) is a subtle innovation that compensates for compositional variation. This subtle change has a profound effect and is essential to achieving logical results since whole genome comparisons based on raw abundance often fail to find a close proximity between closely related species.

Can the same improvement be achieved by dinucleotide relative abundance distance calculations in smaller windows? If so, a delta-scan of one bacterial genome could detect outliers bearing the imprint of a host or foreign signature. Whether this is practical depends on how the variability of the intra-genomic delta-distance grows with shrinking window size. Purely statistical considerations will limit the detectability of small scale deviations. From the investigations of Karlin *et al* [3,4,6,8,9,12,17] it is clear that delta-distances between 50 kb and larger windows show fluctuations that are small compared to distances between closely related species.

Hypothesis formulation

The assessment of bias in dinucleotide relative abundance begins with the "odds ratios" $r_{xy} = f_{xy}/f_x f_y$ where f_x denotes the (normalized) frequency of nucleotide (base) x and f_{xy} is the frequency of dinucleotide (base step) xy in the leading strand. These ratios compare observed dinucleotide frequencies to those expected from the base composition alone under the assumption of statistical independence (i.e., thorough shuffling of the sequence). When r_{xy} is greater (less) than one, xy is over(under)-represented. The symmetrized version r^*_{xy} is computed from frequencies of the sequence concatenated with its inverted complement. The numbers $\{r^*_{xy}\}$ comprise the dinucleotide relative abundance profile [3–6].

The statistical problem is to test the hypotheses that patterns of dinucleotide over- and under-representation in a given genome are invariant. Using symbol f for frequencies in windows, let g be used to represent the global fre-

quencies computed for a complete sequence (a genome or chromosome). The hypothesis asserts that r^*_{xy} in any window is approximately equal to a constant. This constant is the global signature $c^*_{xy} = g^*_{xy}/g^*_x g^*_y$.

Karlin, Landunga, and Blaisdell [8] assessed homogeneity of the dinucleotide relative abundance profile through the delta-distance $\delta^* = (1/16)\ \Sigma\ |r^*_{xy} - c^*_{xy}|$. They provided standards for classifying δ^* in 100 kb windows as follows: "random" ($0 < 1000\delta^* < 15$), "very close" ($15 < 1000\delta^* < 30$), "close" ($30 < 1000\delta^* < 45$), "moderately related" ($45 < 1000\delta^* < 65$), and "distantly related" ($65 < 1000\delta^* < 95$). The upper limit of the "random" range, which typifies thoroughly shuffled sequences, scales as $1/\sqrt{n}$, where n is the number of bases in the window.

The local stability of r_{xy} would be an ancillary result under the assumption of a stationary stochastic process, since then the frequencies converge in probability to fixed limits, $f_x \to p_x$ and $f_{xy}/f_x \to p_{xy}$ as $n \to \infty$. The simplest case would be a homogeneous Markov chain with base step transition probabilities p_{xy} and stationary base composition p_x. In this case the differences $f_x - p_x$ and $f_{xy}/f_x - p_{xy}$ tend to zero as $1/\sqrt{n}$ and it is clear that $r_{xy} - p_{xy}/p_y$ will vanish at the same rate [18]. Moreover, the globally computed quotient $c_{xy} = g_{xy}/g_x g_y$ would be a consistent estimate of p_{xy}/p_y. Thus the separate terms of $\Sigma\ |r_{xy} - c_{xy}|$ tend to zero as $1/\sqrt{n}$ and the same must obviously apply to δ^*.

We start however with the understanding that the sequence is fundamentally nonstationary, exhibiting statistically significant variations in base frequencies between non-overlapping windows [1,2]. Locally or globally estimated Markov models may describe it better than assuming that the bases are independent and identically distributed [19] but they fail to reflect the salient features of natural sequences [20]. For example, Robin and Daudin [21] compared the frequency of a specific motif in the genome sequence of *Haemophilus influenzae* to Markovian predictions and found that the observed frequencies were everywhere higher than predicted.

This point aside, the stationary Markov analogy provides a useful benchmark in assessing local stability of the genomic signature. If nucleotide sequences behaved like Markov chains, then $\delta^*\sqrt{n}$ would not depend on n. We will examine this *scaled delta-distance* $\delta^*\sqrt{n}$ to see if it exhibits any trend. A decreasing ("super-Markov") trend would imply that signature stability emerges as the scale increases and could indicate the breakdown of stability for some window size below 50 kb.

Table 1: Sequences included in this survey

SN	*Sequence*	*Abbr*	*kb*	*GenBank*	*date*
1	*Archaeoglobus fulgidus*	*Aful*	2178	NC_000917	04 Jan 01
2	*Bacillus subtilis*	*Bsub*	4214	NC_000964	12 Oct 01
3	*Borrelia burgdorferi*	*Bbur*	910	AE000783	09 Jan 01
4	*Campylobacter jejuni*	*Cjej*	1641	AL111168	08 Jul 01
5	*Chlamydia pneumoniae J138*	*Cpne*	1226	BA000008	08 Dec 00
6	*Chlamydia trachomatis*	*Ctra*	1042	AE001273	09 Jan 01
7	*Escherichia coli K12*	*Ecol*	4639	U00096	22 Dec 99
8	*Haemophilus influenzae*	*Hinf*	1830	L42023	22 Dec 99
9	*Helicobacter pylori J99*	*Hpyl*	1643	AE001439	09 Jan 01
10	*Methanococcus jannaschii*	*Mjan*	1664	L77117	22 Dec 99
11	*Mycoplasma genitalium*	*Mgen*	580	NC_000908	12 Mar 01
12	*Mycoplasma pneumoniae*	*Mpne*	816	NC_000912	13 Jul 01
13	*Plasmodium falciparum, chr II*	*Pfa2*	947	NC_000910	08 Nov 01
14	*Plasmodium falciparum, chr III*	*Pfa3*	1060	NC_000521	08 Nov 01
15	*Saccharomyces cerevisiae, chr XI*	*Sc11*	666	NC_001143	06 Jun 01
16	*Saccharomyces cerevisiae, chr XV*	*Sc15*	1091	NC_001147	22 Mar 01
17	*Staphylococcus aureus N315*	*Saur*	2813	NC_002745	04 Oct 01
18	*Synechocystis PCC6803*	*Syne*	3573	NC_000911	23 Oct 01
19	*Vibrio cholerae, chromosome I*	*Vch1*	2961	AE003852	09 Jan 01
20	*Vibrio cholerae, chromosome II*	*Vch2*	1072	NC_002506	13 Sep 01
21	*Arabidopsis thaliana, chr IV (1st half)*	*Ath4*	*8750*	NC_003075	21 Aug 01
22	*Human, chromosome XXII*	*Hs22*	*7657*	NT_001039	01 Dec 00
	sum		*52973*		

Scope of the investigation

This survey examines 22 sequences from 19 species and 17 genera. The sequences are listed in Table 1 along with serial numbers (*SN*) and 4-letter abbreviations (*Abbr*). Most of them have been previously studied and found to show stability of the genomic signature in 50 kb windows [9]. The shortest is the 580 kb complete genome of *Mycoplasma genitalium*. The longest complete sequence is the 7657 kb *human chromosome XXII*.

The selected sequences, which are not always complete genomes, fall into two main classes, being typically (1) the chromosome that constitutes the largest single element in a prokaryotic genome or (2) one of the chromosomes in a eukaryotic genome. In the first case, for example, is the *Borrelia burgdorferi* sequence that excludes 21 identified plasmids. The second case is exemplified by *Plasmodium falciparum* where chromosomes II and III are selected to the exclusion of the other 12. Since our investigation focuses on scaling properties of the genomic signature, it is appropriate to consider sequences comprised of many 50 kb contigs, assuming that variation between such contigs is small compared to variation between species.

The present sample, which spans a wide range of G+C proportion, is hoped diverse enough that any consistent trends and features in the statistical picture it produces cannot be easily attributed to chance. A broader range of sequences, including mitochondrial and large viral genomes, has been surveyed by Karlin *et al* [3,4,6,8,9,12,17] using 50 kb and larger windows. This investigation, which applies similar methodology to smaller window sizes, concerns the intra-genomic homogeneity of dinucleotide relative abundance, and inter-sequence distance calculations are beyond its scope.

Our use of nonoverlapping windows is consistent with the methodology employed in prior studies using 50 kb and longer windows. (Overlapping windows with a high percentage overlap would be required to localize and sort out the significance of nonconforming segments that could possibly reflect a foreign signature.) Overlap would introduce statistical dependence between successive windows and such dependence could only complicate the analysis of variance within sequences. Window size was varied by factors of approximately two, the specific values being 50 kb, 25 kb, 10 kb, 5 kb, 2 kb, 1 kb, 500 b, 250 b, and 125 b.

Results and discussion

Increasing trend in mean scaled delta-distance

Table 2 shows mean scaled delta-distance by sequence (Abbr) and window size (n). Scaled delta-distance, defined above as $\delta^*\sqrt{n}$, is a statistical invariant for any benchmark sequence generated by a Markov chain exhibiting the same signature as the given sequence. Except for the Borrelia burgdorferi sequence, the mean scaled delta-distance is increasing in window size for every sequence. The lone exception is a drop, for B. burgdorferi, from 3.580 to 3.553, as window size increases from 10 kb to 25 kb.

The essentially increasing trend in $\delta^*\sqrt{n}$ has an obvious implication for the scalability of standard binning levels used in classifying intra-genomic delta-distance [8]. These levels cannot be re-scaled by the reciprocal square root of window size without admitting that the profiles seen through smaller windows are statistically closer to the global signature. The profiles seen through larger windows obviously tend toward the signature but local fluctuations tend to zero slower than $1/\sqrt{n}$ (i.e., the convergence rate is "sub-Markov").

Quasi-stable hierarchy of mean scaled delta-distances

The rows of Table 2 are grouped as archaea (Aful, Mjan), protist (Pfa2, Pfa3), yeast (Sc11, Sc15), plant (Ath4), human (Hs22), or bacteria (all 14 others). Rows are averaged in groups and the results are plotted in Figure 1. Granted that the curves for archaea and bacteria are almost indistinguishable, we see that ranking mean scaled delta distances by groups produces about the same result for every window size. The one clear exception is the crossing of the curves for plant (Arabidopsis thaliana, chromosome IV) and human (chromosome XXII) between n = 2 kb and n = 5 kb. Thus the view through 50 kb windows shows that the imprint of the global signature is weaker in the latter; but 2 kb and smaller windows show relatively clearer imprints in the human chromosome.

Normality and homoscedasticity of log delta

The increasing trend in all the curves of Figure 1 suggests that local deviations from the global signature are better described by a multiplicative error process (instead of additive). Therefore we examine the (natural) logarithms of the delta-distances in windows and find that they are close to normally distributed with nearly constant variance over the range of window size. Mean values of $-\log(\delta^*)$ are shown in Table 3 with corresponding standard deviations in Table 4.

Each pair of corresponding cells in Tables 3 and 4 defines a normal distribution with mean a variance determined by a sample of observed delta-distances. Each sample is subjected to the Kolmogorov-Smirnov test of fit to the corresponding normal distribution [22]. (We generate a random sample of the same size as the observed sample from a normal distribution with the same parameters. The Kolmogorov-Smirnov two-sample test is then invoked. The null hypothesis is that the observed and simulated samples share a common underlying distribution.) P-values

Table 2: Mean scaled delta-distance by sequence and window size

Abbr	125	250	500	1 k	2 k	5 k	10 k	25 k	50 k
Aful	2.064	2.159	2.304	2.509	2.767	3.120	3.346	3.439	3.619
Bsub	2.133	2.210	2.350	2.570	2.848	3.379	3.984	5.022	6.099
Bbur	2.531	2.609	2.746	2.914	3.123	3.443	3.580	3.553	3.934
Cjej	2.452	2.530	2.695	2.937	3.208	3.551	3.732	3.981	4.047
Cpne	2.162	2.228	2.331	2.487	2.689	3.074	3.316	3.562	3.843
Ctra	2.189	2.259	2.378	2.532	2.724	3.028	3.202	3.356	3.699
Ecol	2.060	2.137	2.259	2.451	2.706	3.105	3.412	3.747	4.118
Hinf	2.209	2.310	2.471	2.671	2.953	3.344	3.685	3.832	4.364
Hpyl	2.256	2.340	2.500	2.740	3.024	3.414	3.809	4.171	4.565
Mjan	2.368	2.474	2.664	2.916	3.232	3.620	4.082	4.634	5.376
Mgen	2.371	2.474	2.644	2.884	3.221	3.753	4.500	5.705	6.761
Mpne	2.247	2.389	2.583	2.892	3.266	3.896	4.273	4.799	5.401
Pfa2	4.352	4.530	4.884	5.384	5.996	6.993	7.655	9.008	10.404
Pfa3	4.265	4.486	4.910	5.459	6.185	7.308	8.549	10.649	11.908
Sc11	2.260	2.344	2.442	2.568	2.717	2.965	3.009	3.226	3.627
Sc15	2.256	2.336	2.451	2.564	2.676	2.828	2.983	3.412	3.681
Saur	2.385	2.480	2.663	2.910	3.222	3.730	4.310	5.201	5.593
Syne	2.036	2.106	2.229	2.433	2.648	2.958	3.173	3.447	3.635
Vch1	2.100	2.192	2.338	2.535	2.801	3.286	3.701	4.231	4.504
Vch2	2.083	2.160	2.271	2.440	2.662	3.116	3.525	4.077	4.728
Ath4	2.581	2.786	3.073	3.434	3.842	4.417	4.934	5.675	6.299
Hs22	2.398	2.581	2.863	3.243	3.723	4.557	5.331	6.627	7.790
average	2.441	2.547	2.728	2.972	3.279	3.763	4.181	4.783	5.357

Table 3: Mean log delta-distance (times -1) by sequence and window size together with log-linear regression results

Abbr	125	250	500	1 k	2 k	5 k	10 k	25 k	50 k	inter	slope	MAR
Aful	1.74	2.04	2.33	2.60	2.85	3.20	3.47	3.89	4.18	0.179	-.400	0.021
Bsub	1.71	2.02	2.31	2.58	2.84	3.15	3.35	3.59	3.77	-.161	-.342	0.059
Bbur	1.55	1.86	2.16	2.45	2.73	3.10	3.41	3.91	4.16	0.560	-.435	0.025
Cjej	1.58	1.89	2.18	2.44	2.71	3.08	3.39	3.78	4.12	0.433	-.417	0.020
Cpne	1.70	2.01	2.32	2.60	2.87	3.20	3.50	3.90	4.21	0.262	-.411	0.021
Ctra	1.68	2.00	2.30	2.58	2.86	3.22	3.50	3.91	4.13	0.267	-.410	0.017
Ecol	1.74	2.05	2.35	2.62	2.88	3.20	3.46	3.82	4.06	0.052	-.382	0.019
Hinf	1.68	1.98	2.27	2.54	2.80	3.15	3.40	3.81	3.99	0.162	-.388	0.021
Hpyl	1.66	1.97	2.26	2.52	2.78	3.13	3.37	3.77	4.06	0.213	-.393	0.017
Mjan	1.61	1.92	2.19	2.46	2.71	3.07	3.30	3.63	3.81	0.112	-.368	0.028
Mgen	1.61	1.91	2.20	2.46	2.70	3.02	3.19	3.41	3.63	-.105	-.332	0.058
Mpne	1.66	1.95	2.22	2.45	2.68	2.97	3.21	3.56	3.76	-.032	-.347	0.022
Pfa2	1.06	1.35	1.63	1.88	2.12	2.43	2.69	3.02	3.18	0.604	-.355	0.027
Pfa3	1.07	1.36	1.63	1.88	2.11	2.43	2.64	2.87	3.12	0.482	-.336	0.038
Sc11	1.66	1.97	2.28	2.59	2.88	3.25	3.58	3.98	4.20	0.389	-.428	0.019
Sc15	1.66	1.97	2.27	2.58	2.89	3.30	3.57	3.93	4.19	0.370	-.426	0.022
Saur	1.60	1.91	2.19	2.46	2.71	3.05	3.26	3.53	3.79	0.066	-.360	0.037
Syne	1.76	2.07	2.37	2.64	2.91	3.28	3.56	3.93	4.23	0.177	-.406	0.012
Vch1	1.72	2.03	2.32	2.59	2.85	3.16	3.40	3.74	3.99	0.023	-.373	0.023
Vch2	1.73	2.04	2.34	2.62	2.88	3.18	3.43	3.74	3.97	-.023	-.369	0.035
Ath4	1.53	1.80	2.05	2.29	2.53	2.85	3.10	3.43	3.68	0.166	-.355	0.008
Hs22	1.62	1.91	2.16	2.39	2.61	2.87	3.05	3.29	3.46	-.250	-.303	0.043

Table 4: Standard deviations of log delta-distance by sequence and window size

Abbr	125	250	500	1 k	2 k	5 k	10 k	25 k	50 k
Aful	.328	.335	.345	.358	.366	.396	.374	.357	.328
Bsub	.331	.338	.350	.377	.398	.440	.491	.500	.545
Bbur	.356	.350	.356	.361	.363	.383	.392	.474	.474
Cjej	.352	.349	.356	.367	.376	.412	.422	.431	.444
Cpne	.334	.335	.345	.362	.350	.374	.439	.478	.579
Ctra	.328	.337	.341	.339	.348	.370	.360	.359	.261
Ecol	.327	.333	.343	.360	.377	.388	.409	.380	.362
Hinf	.340	.345	.357	.372	.400	.424	.438	.477	.324
Hpyl	.344	.354	.363	.382	.402	.429	.415	.474	.518
Mjan	.355	.355	.362	.384	.401	.424	.435	.423	.394
Mgen	.339	.345	.365	.370	.384	.408	.444	.448	.539
Mpne	.344	.342	.348	.344	.367	.380	.337	.395	.272
Pfa2	.466	.452	.448	.451	.455	.479	.447	.500	.422
Pfa3	.454	.451	.463	.486	.482	.519	.534	.499	.523
Sc11	.345	.358	.358	.396	.399	.393	.384	.425	.420
Sc15	.341	.350	.341	.374	.398	.407	.335	.457	.421
Saur	.339	.344	.361	.384	.397	.450	.460	.482	.449
Syne	.336	.346	.370	.387	.407	.433	.440	.415	.438
Vch1	.331	.338	.350	.369	.396	.418	.440	.480	.413
Vch2	.323	.333	.329	.351	.348	.347	.408	.390	.470
Ath4	.364	.370	.374	.374	.379	.409	.420	.453	.479
Hs22	.396	.412	.432	.449	.461	.468	.475	.464	.438

Table 5: P-values from the Kolmogorov-Smirnov test of the normality of log delta-distance by sequence and window size

Abbr	125	250	500	1 k	2 k	5 k	10 k	25 k	50 k
Aful	.0000	.0000	.0046	.1206	.7380	.0733	.9525	.9859	.6249
Bsub	.0000	.0000	.0003	.3529	.3773	.0619	.1773	.2425	.0613
Bbur	.0029	.0446	.0971	.3117	.2172	.2034	.4773	.9762	.9631
Cjej	.0001	.0068	.0263	.4755	.6395	.5143	.3476	.9908	.6351
Cpne	.0000	.0000	.0440	.0092	.7128	.9644	.2337	.0503	.6601
Ctra	.0003	.0000	.0834	.8044	.7337	.4133	.1112	.4046	.9831
Ecol	.0000	.0002	.0070	.3949	.4570	.6386	.6226	.7668	.6447
Hinf	.0000	.0003	.0290	.4296	.6943	.0066	.3318	.6128	.8745
Hpyl	.0008	.0138	.3782	.6892	.6623	.0634	.0718	.6847	.0791
Mjan	.0154	.0000	.0810	.7487	.7469	.2369	.5795	.5344	.4135
Mgen	.0056	.0199	.0783	.5851	.3056	.9500	.9789	.9901	.4792
Mpne	.0004	.1286	.3282	.6395	.4303	.3476	.5625	.4337	.7164
Pfa2	.0529	.0897	.2247	.0913	.1855	.6736	.9891	.8849	.2754
Pfa3	.1893	.0387	.0899	.2086	.2014	.0007	.1906	.1123	.3650
Sc11	.0095	.1440	.0530	.3464	.6546	.2988	.1489	.5010	.5882
Sc15	.0162	.0036	.6316	.8316	.5645	.3060	.9479	1.0000	.3517
Saur	.0000	.0006	.0493	.1764	.1927	.1725	.0321	.7851	.9031
Syne	.0000	.0287	.0137	.0189	.0530	.2909	.0426	.2127	.2563
Vch1	.0000	.0001	.1369	.1247	.4672	.0545	.8498	.3087	.0019
Vch2	.0000	.0000	.2251	.6515	.8533	.8367	.9399	.7912	.6028
Ath4	.1380	.0000	.0047	.1124	.4461	.5067	.8097	.3982	.4813
Hs22	.0000	.0000	.0000	.0000	.0000	.0089	.1619	.0015	.0360
accept	2	3	12	19	21	19	20	21	20

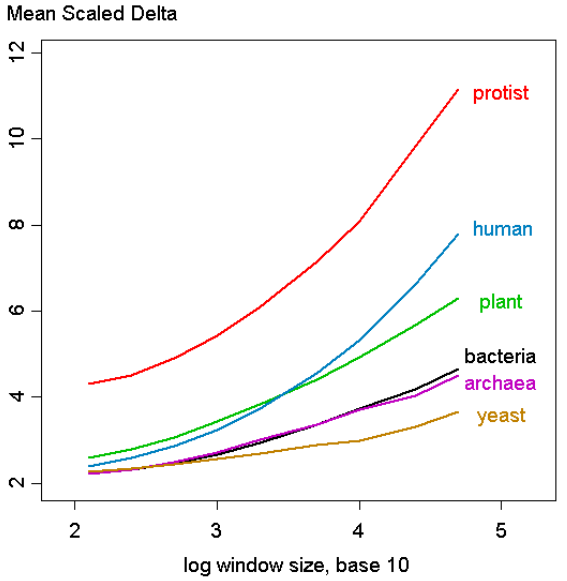

Figure 1
Mean scaled delta-distance versus window size for six subsets of the sequences. The mean intra-genomic delta-distance for windows of a given size on each sequence was computed and scaled up by the square root of window size to give the numbers shown in Table 1. Then the table is condensed by grouping its 22 rows into 6 subsets as defined in the text. Each curve shows the group-wise average.

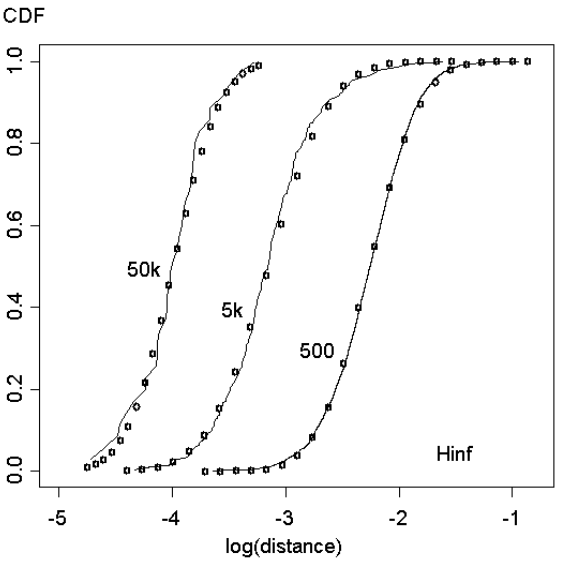

Figure 2
Cumulative distribution functions (CDFs) of log delta-distance computed from windows of three sizes on the*Haemophilus influenzae* genome. The CDF gives the fraction of windows in which log delta-distance did not exceed the abscissa. The empirical CDFs are drawn with connected line segments and they increase in smoothness with the total number of windows. The discrete points (squares) trace the CDFs of corresponding normal distributions having the same mean and variance as the three sets of observations.

from the test, which measure support for the null hypothesis of normality, are shown in Table 5. The bottom row of Table 5 counts (by window size) the number of times the test accepts normality of log(δ*) at the 5% significance level.

Delta-distances tend to fit the log-normal distribution better for window sizes 1 kb and larger. Rejection of log-normality for the smallest window sizes is to some extent a consequence of increasing sample size. The apparent goodness-of-fit actually gets better as window size decreases as illustrated in Figure 2 for three indicated window sizes on the *Haemophilus influenzae* sequence. (The sample CDF approaches a smooth curve as window size declines and the number of windows increases. The heavy plotted points are from log-normal distribution functions with means and variances of the corresponding samples.)

Sub-Markov convergence rate of mean log delta

Since $\delta^*\sqrt{n}$ is essentially increasing in n, we infer that the locally computed dinucleotide relative abundance converges to the global signature more slowly than the reciprocal square root of window size. The log-normality of the intragenomic delta distance suggests a log-linear model of the form $\log(\delta^*) = \alpha - \beta\log(n)$ for intercept (α) and slope (β) coefficients that can be estimated for each sequence (or group) by simple linear regression. Table 3, columns 10 and 11, give intercept and slope coefficients of least squares fits to the scatter plots of mean log(δ*) versus log(n). Column 12 gives the mean absolute residual (MAR) to indicate the closeness of the straight line fit. All estimated slope coefficients are greater than the benchmark value -1/2 implied by the Markovian analogy. Figure 3 shows least squares fits to log-log scatter plots obtained by grouping rows of Table 3 (in the same as way described above for going from Table 2 to Figure 1).

Weakly consistent patterns of intragenomic variability in delta-distance

Residual accumulated delta-distance (RADD) is defined as window size times the cumulative sum of terms $\delta^*(t)$ - mean(δ*), $t = 1, 2, \ldots, T$ where t is the position index (counted in windows from the 5' end) and T is the total number of windows. Here mean(δ*) is just the average of unscaled delta-distances in windows of size n. Since $\delta^*(1) + \ldots + \delta^*(T) = T\,\text{mean}(\delta^*)$, a plot of RADD versus t always returns to zero. The RADD plot will superficially resemble

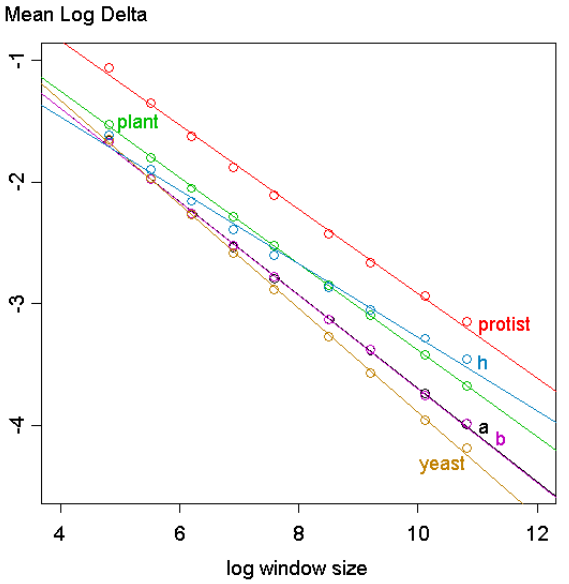

Figure 3
Mean log delta-distance versus window size for six subsets of the sequences. The lines show simple linear regressions of the mean log delta-distance on log window size. The plotted points were obtained by grouping rows of Table 3 as described in the text.

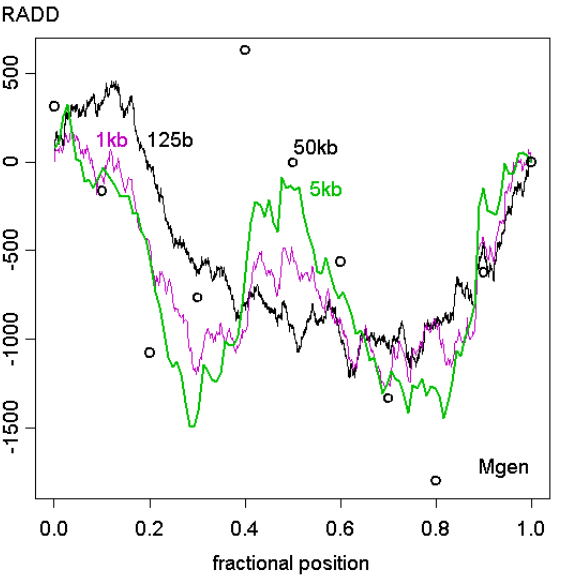

Figure 4
Residual accumulated delta-distance (RADD) versus position for the *Mycoplasma genitalium* sequence. The RADD is plotted against fractional position for each of four indicated window sizes. The full horizontal scale is 580 kb.

random walks obtained by integrating counts of purine minus pyrimidine bases. Such "walking plots" are useful in depicting long range compositional biases concomitant to replication [7] and their self-similarity with respect to re-scaling has been said to instantiate the "fractal geometry of nature [23,24]."

Here we use the RADD plot to examine how well delta-distances in smaller windows are smoothed by corresponding larger windows. For example, take a single 50 kb window on the *M. genitalium* sequence, compute its delta-distance from the global signature, then divide it into 50 contiguous 1 kb windows and repeat the calculation for each piece. Do the 50 delta-distances exhibit random fluctuations about the level indicated by the single larger window? If so, the patterns would be judged consistent, and that result would typify a stationary process. Such consistency is only weakly evidenced In Figure 4, however, as larger scales give rise to emergent trends. The same effect is seen in Figure 5 where the 125 b and 1 kb traces track each other closely but diverge from the other traces in the central region.

Conclusions

The imprint of the global signature is locally pervasive on all scales considered in the sequences that were scanned. No lower bound can yet be placed on the local scale on

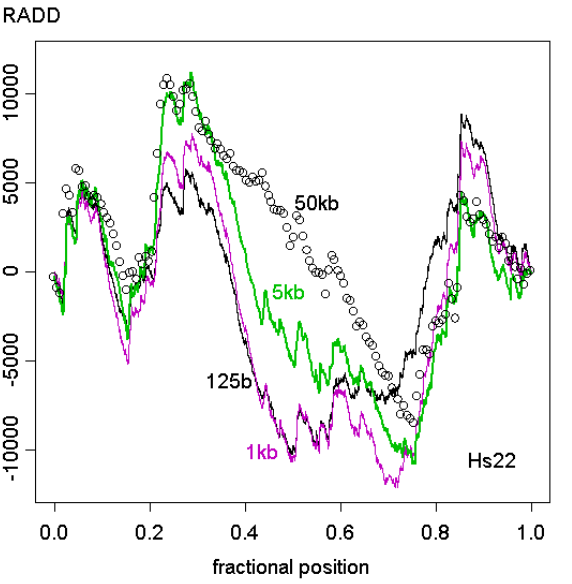

Figure 5
Residual accumulated delta-distance (RADD) versus position for the *Human chromosome XXII* sequence. The RADD is plotted against fractional position for each of four indicated window sizes. The full horizontal scale is 7650 kb.

which the global signature is reflected. The inter-genomic hierarchy of mean intra-genomic delta-distances is essentially preserved across the range of window sizes. Intra-genomic delta-distance is approximately log-normally distributed (in windows down to 1 kb) and the variance of log-delta is fairly uniform across the set of sequences. Delta-distance tends to zero with increasing window size but the rate of this convergence is significantly slower than for simple random processes.

Methods

The sequences listed in Table 1 were downloaded from the (National Center for Biotechnology Information) GenBank [25] in FASTA format and saved as plain text files. The last two columns of Table 1 provide GenBank accession numbers and approximate dates of the revisions used in this analysis. The sequences, as text files, were processed by routines written in SPlus, Version 4.5. The texts were read in blocks (contigs) of 50 kb (when computing delta-distances in 50 kb and 25 kb windows) or 20 kb (for smaller windows). Counts of overlapping base steps in windows were computed with the last base of one window serving as the first base of the next. Thus the number of base steps in a window is equal to the number of bases (not one less). However, one base step was uncounted at the start of every block. Incomplete windows in the last block were discarded and blocks past end-of-file were ignored. Sometimes complete blocks near end-of-file were omitted. With the exception of *Arabidopsis thaliana, chromosome IV*, the total sequence lengths in kilobases are listed in the fourth column (*kb*) of Table 1. All corresponding sample lengths are at least 96% of total length, and most are close to 99%, when texts were read in 20 kb blocks. For *A. thaliana*, however, the global signature was computed from a scan of the 99% complete sequence; but local delta-distances from the global signature were computed only for the first 50% of the complete sequence due to computing difficulties. The 8750 kb length for this sequence in Table 1 is therefore about half the total bases in the chromosome.

Authors' contributions

RHB conceived the study and carried out the computations. RWJ participated in experimental design and data analysis. Both authors read and approved the final manuscript.

List of abbreviations

CDF: Cumulative Distribution Function

MAR: Mean Absolute Residual

RADD: Residual Accumulated Delta-Distance

Sequence abbreviations are as shown in Table 1.

References

1. Karlin S, Brendel V: **Patchiness and correlations in DNA sequences.** *Science* 1993, **259:**667-679
2. Braun JV, Müller H-G: **Statistical methods for DNA sequence segmentation.** *Statistical Science* 1998, **13:**142-162 [projecteuclid.org/Dienst/UI/1.0/Display/euclid.ss/1028905933?abstract]
3. Mrázek J, Karlin S: **Strand compositional asymmetry in bacterial and large viral genomes.** *Proc Natl Acad Sci USA* 1998, **95:**3720-3725
4. Karlin S, Mrázek J, Campbell AM: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriology* 1997, **179:**3899-3913
5. Russell GJ, Subak-Sharpe JH: **Similarity of the general designs of protochordates and invertebrates.** *Nature* 1977, **266:**533-535
6. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends in Genetics* 1995, **11:**283-290
7. Freeman JM, Plasterer TN, Smith TF, Mohr SC: **Patterns of genome organization in bacteria.** *Science* 1996, **279:**1827-1829
8. Karlin S, Landunga I, Blaisdell BE: **Heterogeneity of genomes: measures and values.** *Proc Natl Acad Sci USA* 1994, **91:**12837-12841
9. Campbell A, Mrázek J, Karlin S: **Genome signature comparisons among prokaryote, plasmid, and mitochondiral DNA.** *Proc Natl Acad Sci USA* 1999, **96:**9184-9189
10. Frick DN, Richardson CC: **DNA primases.** *Annu Rev Biochem* 2001, **70:**39-80
11. Arnold J, Cuticchia AJ, Newsome DA, Jennings WW, Ivarie R: **Mono-through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis.** *Nucleic Acids Res* 1988, **18:**7145-7158
12. Karlin S, Landunga I: **Comparisons of eukaryotic genome sequences.** *Proc Natl Acad Sci USA* 1994, **91:**12832-12836
13. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304
14. Blaisdell BE: **Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences.** *J Molec Evol* 1989, **29:**526-537
15. Pevzner PA: **Statistical distance between texts and filtration methods in sequence comparison.** *ABIOS* 1992, **8:**121-127
16. Gibbs AJ, Dale MB, Kinns HR, MacKenzie HG: **The transition matrix method for comparing sequences.** *Systematic Zoology* 1971, **20:**417-425
17. Cardon LR, Burge C, Cayton DA, Karlin S: **Pervasive CpG suppression in animal and mitochondrial genomes.** *Proc Natl Acad Sci USA* 1994, **91:**3799-3803
18. Billingsley P: **Statistical methods in Markov chains.** *Ann Math Stat* 1961, **12:**488-497
19. Avery PJ, Henderson DA: **Fitting Markov chain models to discrete state series such as DNA sequences.** *Applied Statistics* 1999, **48:**53-61
20. Pevzner PA: **Nucleotide sequences versus Markov models.** *Computers Chem* 1992, **16:**103-106
21. Robin S, Daudin J-J: **Exact distribution of the distances between any occurrences of a set of words.** *Ann Inst Statist Math* 2001, **4:**895-905
22. Daniel WW: **Applied Nonparametric Statistics,** *Boston, PWS-Kent Pub Co* 1990
23. Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simon M, Stanley HE: **Finite-size effects on long-range correlations: implications for analyzing DNA sequences.** *Phys Rev E* 1993, **47:**3730-3733
24. Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simon M, Stanley HE: **Statistical properties of DNA sequences.** *Physica A* 1995, **221:**180-192
25. Benson DA, I Karsch-Mizrachi, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2000, **28:**15-18