

DINUCLEOTIDE DENSITY FLUCTUATIONS IN GENOME SEQUENCES

R.H. BARAN

Genostat, 4508 Cheltenham Drive, Bethesda, MD 20814

email: baran@genostat.com

DNA sequences exhibit significant intra-genomic fluctuations with respect to a model of conditional uniformity under which (1) the expected frequency of each dinucleotide (base step) is proportional to the product of the marginal (base) frequencies and (2) the observed number follows a Poisson distribution with that conditional mean. The empirical relation between dinucleotide relative density (DRD) and Shannon mutual information in base steps is explained by broadening the model to include local modulation of the conditional mean. This modulation takes the form of a power law as local and global DRD components are log-linearly related.

1. Introduction. Karlin and Brendel (1993) posed the question, "Can DNA sequence variation be reasonably modeled by stochastic processes of a tractable genre?" A viable model will need to account for the meaningful statistical invariants that have been discovered so far, including the dinucleotide relative density profile, which is consistent within genomes (Karlin, Landunga, and Blaisdell, 1994; Karlin and Mrazek, 1997; Campbell, Mrazek, and Karlin, 1999), and the Shannon mutual information in proximate base pairs, which varies little between genomes when coding and noncoding sequences are treated separately (Grosse, Herzel, Buldyrev, and Stanley, 2000). The dinucleotide relative density (DRD) and the Shannon mutual information (SMI) in ungapped base pairs are mathematically related in a way that can be rigorously described in the context of stationary random processes such as simple Markov chains that sometimes find use in modeling nucleotide sequences when the statistical independence of successive bases is too crude of an approximation (Pevzner, 1992; Avery and Henderson, 1999; Robin and Daudin, 2001). But such models are incompatible with the compositional heterogeneity that is found on all scales in every sequence.

In fields ranging from microbiology to cosmology, the study of density fluctuations usually begins with the hypothesis that the number of objects (microbes, stars, etc.) in a "large enough" region has a statistically uniform (Poisson) distribution. This hypothesis is adapted to the analysis of DRD fluctuations in sufficiently long (contiguous) segments of the genome sequence by postulating that (1) the expected frequency of dinucleotide (base step) ij is everywhere proportional to the product of base frequencies i and j and that (2) the observed dinucleotide frequency follows a Poisson distribution with the given conditional mean. The proportionality factors are the 16 components of the DRD profile computed (globally) from the base and base step counts of the whole

sequence. Thus the frequency of any base step is modulated by the local base composition but the relative density remains close to the global value. Under this conditional Poisson model, local fluctuations in the SMI are described by the usual chi-squared theory, and local variation in DRD can be studied by computer simulation using the actual base compositions of the segments.

The focal point of this investigation is a scatter plot of local variation in DRD and co-located estimates of SMI in segments spanning the complete sequence of a chromosome. For the 26 complete sequences examined there, with segment sizes of 1, 5, and 25 kilobases (kb), such plots exhibit a characteristic pattern that can be explained by an adjustment to the model under which the DRD components vary locally within the sequence subject to their transformation by a common exponent. This exponential transformation corresponds to a rotation in the log-log plot of local versus global DRD components and implies a linear relationship between the exponent and SMI. While the adjusted model neither explains the compositional heterogeneity of the sequence nor defines a conditionally stationary stochastic process (that develops one base step at a time), its three defining properties are necessary attributes of any viable stochastic model.

2. Definitions and notations. If f_{ij} denotes the frequency of ij dinucleotides in some n -long segment starting at a given location, and f_j is the frequency of j nucleotides, then the quotient $r_{ij} = f_{ij}/(f_i f_j)$ is the *relative density* of dinucleotide ij in the segment. (Note that overlapping dinucleotides are counted. E.g., the 4-long sequence ACGT gives $f_j = 1/4$ for each j and $f_{ij} = 1/3$ for $ij = AC, CG,$ and GT ; but f_{ij} is zero otherwise.) These sixteen quotients are collected in a vector or square matrix that constitutes the DRD profile. Every DRD component tends toward a constant limit irrespective of location as n increases. For global frequencies denoted and g_{ij} and g_j , let $\rho_{ij} = g_{ij}/(g_i g_j)$. This global value, obtained by scanning the whole sequence from left to right (5' to 3'), is regarded as a consistent estimator of a corresponding genomic constant. Averaging the absolute values of the local fluctuations in the sixteen components we obtain

$$(1) \quad \delta = (1/16)\sum\sum|r_{ij} - \rho_{ij}|$$

which, except for symmetrization (equivalent to concatenation of the sequence with its inverted complement), is the measure introduced by Karlin *et al.* to assess variability within genomes. The local variation δ tends to be log-normally distributed and its average approaches zero slightly slower than $n^{-1/2}$ (Jernigan and Baran, 2002).

The SMI can be regarded as a measure of association between two random variables as noted by Román-Roldán *et al.* (1996). It is computed as the average value of the logarithm of the reciprocal likelihood ratio for testing the hypothesis of mutual independence (Yuan and Clarke, 1999). The reciprocal likelihood ratio is presently the joint probability of i and j divided by the product of their marginal probabilities. Substituting relative frequencies for the probabilities, with f_{ij} and f_i for the joint and marginal, respectively, this ratio is the DRD. Hence the SMI is estimated at a given location (framed by a window of some fixed size n) as the average value of $\log(r_{ij})$. The likelihood ratio

statistic for testing the hypothesis of independence, $G^2 = 2n \sum \sum f_{ij} \log(r_{ij})$, which is also called the *deviance* (Agresti, 1990), is asymptotically (for large enough n) distributed as chi-squared with $(m-1)^2$ degrees of freedom when there are m states, categories, or symbols (Avery and Henderson, 1999). With $m = 4$ nucleotides, there are $3^2 = 9$ degrees of freedom, and an unbiased estimator of the noncentrality parameter of this chi-squared distribution is

$$(2) \quad 2I = 2 \sum \sum f_{ij} \log(r_{ij}) - 9/n$$

which is twice the locally estimated SMI. This *unbiased* estimator may be less than zero even though the SMI is non-negative. The SMI approaches zero as the sequence is repeatedly shuffled since then the bases at neighboring positions tend toward statistical independence. Larger values of the SMI imply stronger dependence up to a maximum of $I = \log(m) - (m-1)^2/(2n) < 1.099$ nats per nucleotide (using natural logarithms here and throughout).

3. Hypothesis formulation. The simplest hypothesis of spatial uniformity holds that the count of objects in a region of extent x is a Poisson random number with expected value λx . When m different types of objects are considered, the type-specific counts n_k are Poisson random numbers with respective means $\lambda_k x$ ($k = 1, 2, \dots, m$). When the sum of all m type counts is constrained to equal a fixed sample size n , the type counts are multinomially distributed with expected values $n \lambda_k / \sum \lambda_k$ (Agresti, 1990). In the present problem we have $m^2 = 16$ types as k is replaced with ij and $n f_i f_j = (1/n) n_i n_j$ corresponds to x . Note that n is the length (number of bases in) the segment and n_{ij} is the count of ij base steps. Then the distribution of n_{ij} conditioned on the (mononucleotide) composition of the sample is Poisson with mean value $n \rho_{ij} f_i f_j$. Thus the constraint $\sum \sum (n_{ij}/n) = \sum \sum f_{ij} = 1$ implies that n_{ij} is conditionally multinomial with expectation $n \rho_{ij} f_i f_j / D$ where $D = \sum \sum \rho_{ij} f_i f_j$.

Under these assumptions the deviance statistic G^2 (defined above) is distributed (for large enough n) as chi-squared with noncentrality parameter $2nI$ as written in equation (2); and the global estimator of the SMI is

$$(3) \quad I_g = \sum \sum g_{ij} \log(\rho_{ij}) - 9/(2T)$$

where T is total sequence length. Moreover r_{ij} will tend towards ρ_{ij}/D so that $\delta \approx |1 - 1/D|$ and $I \approx \sum \sum f_{ij} \log(\rho_{ij}) - \log D - 9/(2n)$ under the conditional Poisson model. Since g_{ij} is the average of the f_{ij} in all contiguous segments of the sequence, the average value of the SMI in segments will differ from I_g by an amount that tends to zero as D approaches 1.

In the absence of compositional heterogeneity, f_{ij} and f_j would approximate the global frequencies g_{ij} and g_j with variance on the order of $1/n$ and $D \rightarrow 1$ with variance $O(1/n^2)$. In the presence of compositional heterogeneity, δ and I can be computed by the Monte Carlo method, taking the actual composition $\{f_j\}$ for each segment, and the DRD profile $\rho_{ij} = g_{ij}/(g_i g_j)$ of the whole sequence, to generate multinomial samples. The resulting sequence of simulated data points $(I, \delta)^*$ reflects the distribution of the observations under the

conditional Poisson model. One check on the fidelity of the simulation is to compare I_g from equation (3) to the average I^* since these should nearly coincide according to the argument just stated.

Table 1. Sequences identified by serial number (SN), species name (and strain), chromosome number, size, GenBank accession number, and the approximate date of the revision used.

SN	Species, chromosome	Size (kb)	GenBank	Revised Mo Yr
1	<i>Arabidopsis thaliana</i> , chr. IV	17546.36	NC_003075	Aug 01
2	<i>Archaeoglobus fulgidus</i>	2157.84	NC_000917	Jan 01
3	<i>Bacillus subtilis</i>	4199.92	NC_000964	Oct 01
4	<i>Borrelia burgdorferi</i>	909.66	AE000783	Jan 01
5	<i>Campylobacter jejuni</i>	1638.36	AL111168	Jul 01
6	<i>Chlamydia pneumoniae</i>	1218.78	BA000008	Dec 00
7	<i>Chlamydia trachomatis</i>	1038.96	AE001273	Jan 01
8	<i>Escherichia coli</i> K12	4595.40	U00096	Dec 99
9	<i>Haemophilus influenzae</i> Rd	1827.97	L42023	Dec 99
10	<i>Helicobacter pylori</i> J99	1638.36	AE001439	Jan 01
11	<i>Homo sapiens</i> , chr. XXII	7646.85	NT_001039	Dec 00
12	<i>Methanobacterium thermoautotrophicum</i> Delta H	1708.29	AE000666	Jun 02
13	<i>Methanococcus jannaschii</i>	1658.32	L77117	Dec 99
14	<i>Mycobacterium tuberculosis</i>	4399.06	AE000516	Nov 02
15	<i>Mycoplasma genitalium</i>	579.42	NC_000908	Mar 01
16	<i>Mycoplasma pneumoniae</i>	815.51	NC_000912	Jul 01
17	<i>Plasmodium falciparum</i> , chr. II	946.05	NC_000910	Oct 01
18	<i>Plasmodium falciparum</i> , chr. III	1058.90	NC_000521	Oct 01
19	<i>Saccharomyces cerevisiae</i> , chr. XI	659.34	NC_001143	Jun 01
20	<i>Saccharomyces cerevisiae</i> , chr. XV	1078.92	NC_001147	Mar 01
21	<i>Salmonella enterica</i>	4705.29	NC_003198	Jan 03
22	<i>Staphylococcus aureus</i> N315	2797.20	NC_002745	Oct 01
23	<i>Streptomyces coelicolor</i> A3	8591.40	NC_003888	Mar 03
24	<i>Synechocystis</i> PCC6803	3566.43	NC_000911	Oct 01
25	<i>Vibrio cholerae</i> , chr. I	2956.99	AE003852	Jan 01
26	<i>Vibrio cholerae</i> , chr. II	1058.93	NC_002506	Sep 01

4. Data selection and processing. With analytic routines written in SPlus (Version 4.5), δ and I were computed in contiguous segments of length $n = 1, 5, 10,$ and 25 kb spanning the complete sequences of 26 chromosomes identified in Table 1 together with GenBank accession numbers and the approximate date of the revision that was downloaded. Included in this list are 23 species from 21 genera including 14 bacteria, 3 archaea, yeast, protist, plant, and human. The sum of all the sequence lengths exceeds 80,000 kb. The longest sequence, *Arabidopsis thaliana*, chromosome IV, at 17,550 kb, indicates the upper bound

on what the program can accomplish running overnight on a personal computer. Computing δ requires a preliminary pass through the sequence to count all base steps and thus produce the global estimates denoted ρ_{ij} . The sizes listed in Table 1 are the base step counts. They are less than total sequence length (but always at least 99% of total) because the routine reads sequences in blocks of typically 20 kb and blocks extending beyond end of file are omitted. Also the number of transitions is one less than the block length. The next section will focus mainly on representative results for length $n = 5$ kb.

Table 2. Summary of computed results for the 26 sequences listed by serial number (SN) and 4-letter abbreviation (*Abbr*). The SMI is given as the average over all 5kb segments and as a global value. The average variation of the DRD in segments (a) is listed side-by-side with its expected value based on a single simulation with constant DRD.

SN	<i>Abbr</i>	2000I		1000 δ		% (c) absorbed	1000 δ residual a(1-c)-b
		local	global	(a) observed	(b) expected		
1	<i>Ath4</i>	22.70	17.97	70.42	37.13	33.63	9.61
2	<i>Aful</i>	38.44	36.27	58.17	36.83	30.07	3.84
3	<i>Bsub</i>	36.30	33.60	58.37	37.05	28.76	4.53
4	<i>Bbur</i>	47.05	46.47	70.61	48.00	18.00	9.90
5	<i>Cjej</i>	54.21	51.87	65.92	43.78	30.43	2.08
6	<i>Cpne</i>	29.80	27.03	55.14	35.92	31.53	1.83
7	<i>Ctra</i>	29.99	27.15	55.82	36.77	30.09	2.25
8	<i>Ecol</i>	29.37	25.51	61.64	35.34	36.56	3.77
9	<i>Hinf</i>	37.06	34.30	56.76	36.49	34.70	0.57
10	<i>Hpyl</i>	64.16	61.04	64.50	39.46	28.37	6.74
11	<i>Hs22</i>	84.10	73.75	73.67	40.89	35.71	6.48
12	<i>Mthe</i>	47.45	43.97	58.14	36.70	28.53	4.86
13	<i>Mjan</i>	43.26	41.86	76.85	46.17	20.15	15.19
14	<i>Mtub</i>	27.78	23.97	58.88	37.64	28.97	4.18
15	<i>Mgen</i>	48.26	44.78	68.55	39.64	22.16	13.71
16	<i>Mpne</i>	41.21	37.62	65.94	38.20	23.74	12.09
17	<i>Pfa2</i>	23.77	10.31	114.34	56.49	49.44	1.32
18	<i>Pfa3</i>	26.40	10.15	122.96	57.02	55.61	-2.32
19	<i>Sc11</i>	16.07	14.65	50.41	36.86	26.67	0.11
20	<i>Sc15</i>	15.20	13.82	49.96	35.87	27.14	0.53
21	<i>Sent</i>	31.84	27.50	64.84	35.78	37.30	4.87
22	<i>Saur</i>	16.14	11.85	65.08	39.77	36.98	1.24
23	<i>Scoe</i>	26.90	22.18	69.23	40.51	31.80	6.71
24	<i>Syne</i>	47.99	46.55	51.05	35.57	35.35	-2.57
25	<i>Vch1</i>	28.77	25.06	57.48	34.53	37.79	1.23
26	<i>Vch2</i>	29.55	26.93	53.12	35.12	35.88	-1.07

5. Results and discussion. Table 2 compares the average values of the observed I and δ in 5 kb segments to predictions based on the conditional Poisson model. The first pair of data columns shows, for each complete sequence, the average value of $2000I$ from equation (2) in segments and the corresponding global estimate from equation (3). The global estimate of SMI differs from the average of the simulated values (not shown) by less than 4% in every case with a mean absolute difference of 1%. Percentage difference is computed as $200(I - I^*)/(I + I^*)$. The second pair of columns shows average values of 1000δ for (observed) 5 kb segments of the nucleotide sequence and for (expected) simulated samples of size $n = 5$ kb with the same base compositions as the segments. These column pairs are plotted in Figures 1 and 2. Figure 1 reveals that average SMI in segments equals or exceeds the global estimate. In Figure 2 the plotted points (solid triangles) lie above the dashed line of slope 1.35 showing that the average variation exceeds the model-based prediction by at least 35 per cent in every case. Except for the two chromosomes of *Plasmodium falciparum*, the excess variation is between 35% and 100% of prediction.

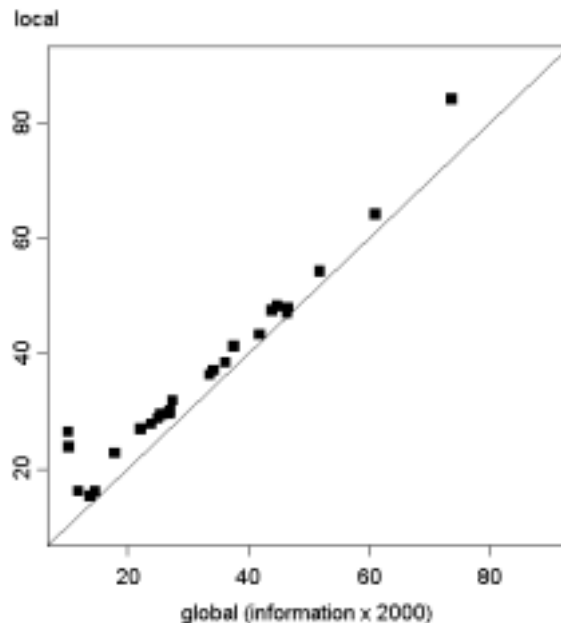


Figure 1. Scatter plot of average local SMI (x2000) versus the corresponding global estimate in 26 chromosome sequences with segment size 5 kb. The plotted points lie above the line of unit slope through the origin and thus the local average exceeds the global value in every case.

The log-normal distribution of the intra-genomic delta distance with symmetrization (Jernigan and Baran, 2002) carries over to the variation in equation (1). As this sum of absolute differences in the DRD components is log-normal, it is not surprising that a weighted sum of the logarithms of the DRD

components is normally distributed. The deviance G^2 defined in connection with equation (2) is such a weighted sum and thus its distribution in 5 kb segments of

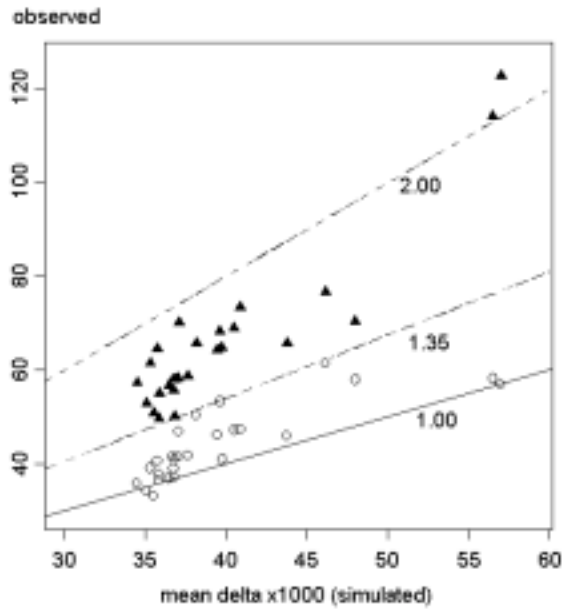


Figure 2. Scatter plot of observed average local variation from the global DRD profile in 26 chromosome sequences (solid triangles) versus predictions from model-based simulation. Observation exceeds prediction by at least 35%, but generally not more than 100%, as shown by drawing dashed lines of indicated slopes through the origin. The observed variations are reduced by the absorption percentages (Table 2) implied by the adjusted model and re-plotted as open circles. Segment size is 5 kb.

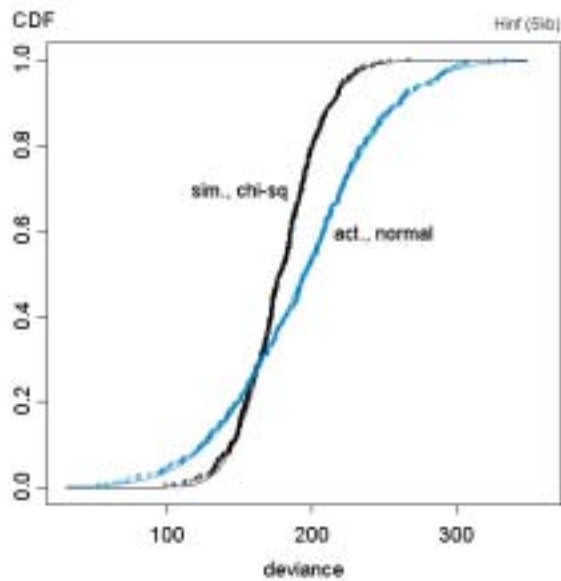


Figure 3. Cumulative distribution function (CDF) of the deviance statistics for 5kb segments of the *H. influenzae* sequence.

the *Haemophilus influenzae* sequence, in Figure 3, is very close to normally distributed. The corresponding deviance statistics from the model-based simulation are indeed distributed as chi-squared with 9 degrees of freedom (with noncentrality parameter 5 times the average value of $2000I$). When observations follow a Poisson distribution at every local site, but the parameter is subject to normal fluctuations from its global mean, the observations (pooled from all the local sites) will be log-normally distributed (Christensen and Waagepetersen, 2002).

The joint distribution of I and δ is studied through scatter plots like those presented for 5 kb segments of the *H. influenzae* sequence in Figure 4 (for the conditional Poisson model) and Figure 5 (actually observed). Both I and δ exhibit greater variance than predicted by the model. Comparing the (horizontal) dispersion of the SMI in Figure 5 to that in Figure 4, the information dispersion ratio is $IDR = \sigma(I)/\sigma(I^*) = 1.87$ in terms of the standard deviations. For the (vertical) dispersion of the local variation, the variation dispersion ratio is $VDR = \sigma(\delta)/\sigma(\delta^*) = 2.45$. These dispersion ratios are very close to the corresponding medians for the whole data set, which are 1.85 and 2.49, respectively, with reference to Table 3 where VDR and IDR are listed for all 26 sequences. The IDR ranges across the data set from a minimum of 1.26 (*Borrelia burgdorferi*) to a maximum of 11.1 (*P. falciparum*, chromosome III). Except for the *Plasmodium* sequences, the largest IDR is 2.86 (*Arabidopsis thaliana*, chromosome IV). The VDR ranges from a minimum of 1.75 (*Saccharomyces cerevisiae*, chromosome XI) to a maximum of 5.31 (*P. falciparum*, chromosome III). Except for the *Plasmodium* sequences, the largest VDR is 3.84 (human chromosome XXII).

5.1 Systematic fluctuations. The excess dispersions of I and δ with respect to model-based predictions may be explained in two ways: Either the conditional Poisson model is fundamentally flawed or the components of the DRD profile exhibit some intrinsic variability. In the latter case we can inquire if there is any system to the excess variation in the DRD components. The simplest systems that could be involved will be described by direct proportionalities like $r_{ij} \propto \rho_{ij}$ and $\log(r_{ij}) \propto \log(\rho_{ij})$ where the same proportionality factor applies to all base steps. The first form seems untenable because DRD components obviously engage in a competitive relationship. E.g., if the frequency of CG is strongly suppressed, relative to the product of the mononucleotide frequencies, then there must be a corresponding enhancement in one or more of the frequencies for CA, CC, and CT. Hence all row and column sums in the 4-by-4 matrix of (global) DRD profile components are in the range 4.0 ± 0.5 while individual components range from 0.313 (for CG in human chromosome XXII) to 1.541 (for GC in *Helicobacter pylori*). The sum of all sixteen components is in the range 16.0 ± 0.5 for all 26 sequences.

Suppose however that the genomic constant ρ_{ij} is modulated locally according to a power law so that the local frequency of base step ij has expected value

$$(4) \quad f_{ij}(\beta) = \frac{\rho_{ij}^{\beta} f_i f_j}{\sum \sum \rho_{ij}^{\beta} f_i f_j}$$

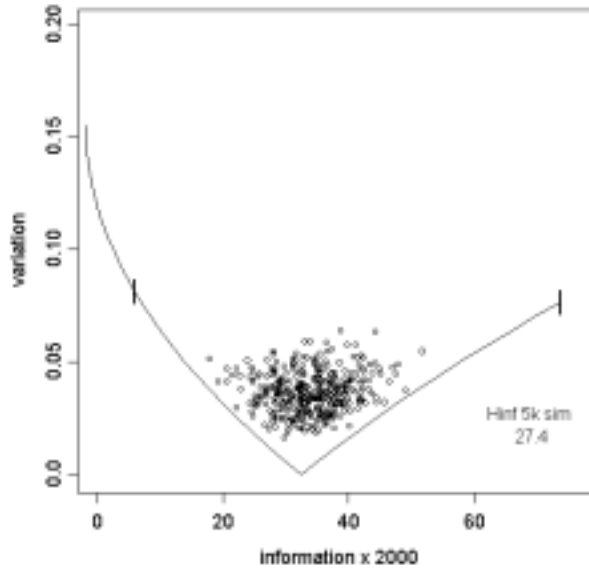


Figure 4. Scatter plot of variation versus information in 5 kb segments of the *H. influenzae* sequence in a simulation based on the conditional Poisson model. Each simulated segment has the actual base composition but dinucleotide frequencies are generated by the model.

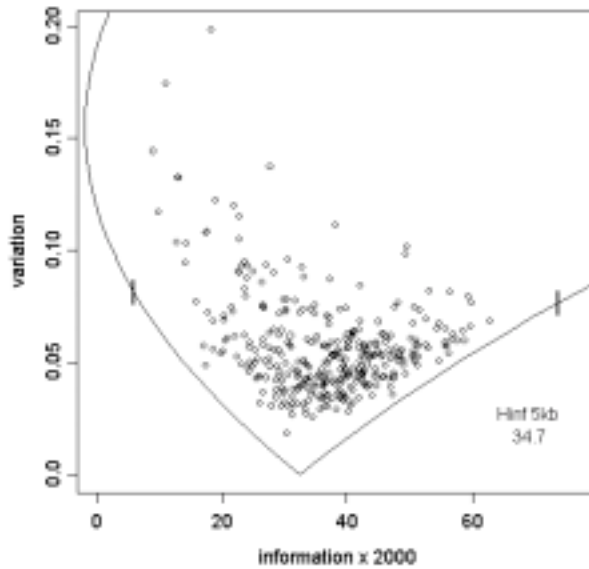


Figure 5. Scatter plot of variation versus information in 5 kb segments of the *H. influenzae* sequence. The fitted curve, which absorbs 34.7% of the total

variation, is drawn by holding base composition and the DRD components at their global average values and varying the exponent, β . The vertical tick marks on the curve indicate $\beta = \pm 1/2$ with respect the minimum at $\beta = 1$.

Table 3. Information (IDR) and variation (VDR) dispersion ratios, odds ratio (OR), P-value (as -1 times the base 10 logarithm), standard deviation of the estimated β , and simple correlation coefficient between I and β (computed with 5% trim).

SN	Abbr	IDR	VDR	OR	$-\lg t(P)$	$\sigma(\beta)$	$\text{cor}(I, \beta)$
1	<i>Ath4</i>	2.868	2.649	8.012	16	0.369	0.882
2	<i>Aful</i>	1.944	2.234	5.923	16	0.168	0.952
3	<i>Bsub</i>	1.824	2.742	8.062	16	0.217	0.947
4	<i>Bbur</i>	1.256	2.140	2.109	5	0.126	0.700
5	<i>Cjej</i>	2.046	2.244	4.216	16	0.139	0.731
6	<i>Cpne</i>	1.732	2.215	4.474	16	0.154	0.895
7	<i>Ctra</i>	1.741	1.958	4.533	15	0.146	0.914
8	<i>Ecol</i>	1.836	2.832	8.470	16	0.188	0.905
9	<i>Hinf</i>	1.874	2.447	8.688	16	0.189	0.951
10	<i>Hpyl</i>	1.777	2.549	6.333	16	0.123	0.946
11	<i>Hs22</i>	2.324	3.840	2.262	16	0.193	0.592
12	<i>Mthe</i>	1.645	2.705	7.812	16	0.127	0.909
13	<i>Mjan</i>	1.354	2.237	3.600	16	0.190	0.810
14	<i>Mtub</i>	2.366	3.319	4.107	16	0.164	0.820
15	<i>Mgen</i>	2.040	2.420	2.913	5	0.183	0.653
16	<i>Mpne</i>	1.633	2.447	3.400	9	0.136	0.858
17	<i>Pfa2</i>	6.953	3.913	2.160	5	0.278	0.326
18	<i>Pfa3</i>	11.133	5.317	2.490	7	0.307	0.432
19	<i>Sc11</i>	1.503	1.753	5.294	11	0.192	0.890
20	<i>Sc15</i>	1.774	1.986	4.516	15	0.208	0.869
21	<i>Sent</i>	2.032	3.155	7.774	16	0.214	0.896
22	<i>Saur</i>	2.847	2.817	5.535	16	0.316	0.839
23	<i>Scoe</i>	1.640	2.353	2.963	16	0.191	0.701
24	<i>Syne</i>	1.784	2.724	7.133	16	0.142	0.973
25	<i>Vchl</i>	2.182	2.915	9.531	16	0.227	0.940
26	<i>Vch2</i>	1.980	1.938	6.640	16	0.174	0.951

for some exponent β . The denominator normalizes the sixteen frequencies. Setting $\beta = 1$ just returns the basic model. Large positive exponents $\beta \gg 1$ will concentrate the distribution in the base steps with the largest relative density. Small positive exponents $\beta \approx 0$ will randomize the sequence as $f_{ij}(0) = f_i f_j$. Dividing equation (4) by the product of the marginals and taking logarithms gives

$$(5) \quad \log r_{ij}(\beta) = \beta \log \rho_{ij} - \log(\sum \sum \rho_{ij}^\beta f_i f_j).$$

The double sum can be denoted $D(\beta)$ consistent with the earlier discussion. When this relation holds, a log-log plot of local versus global DRD components

will exhibit clustering along the line with slope β and intercept $-\log D(\beta)$. Moreover a log-log linear regression of the local DRD components on their global counterparts should give reasonable estimates of the slope and intercept.

5.2. *Rotational trends.* Clustering along the line with slope β implies a general clockwise (or counter-clockwise) rotation of the points $\log(\rho_{ij}, r_{ij})$ when β is less than (or greater than) one. Such behavior is clearly visible in Figure 6, a log-log plot of local versus global DRD components for five contiguous 10 kb segments of the *H. influenzae* sequence (beginning 150 kb from sequence start). Indexing the bases alphabetically, and indexing the base steps as $i + 4(j-1)$, the index of each base step is printed near the top for each abscissa $\log(\rho_{ij})$. Thus the smallest global DRD component (for TA) has index 4 and the largest (for GC) has index 7. The vertical axis is drawn as a dashed line and the line of unit slope ($\beta = 1$) through the origin is solid. Each segment produces sixteen DRD components indicated by a single symbol. The plotted circles, for the first segment, lie below the solid line when left of the vertical axis (with the exception of CT) and above the solid line when right (with the exception of CG). Thus 14 of 16 components show clockwise rotation. The plotted crosses, for the third segment, are rotated counter-clockwise (with the exceptions of AT, CC, and GG, whose relative densities are closest to 1).

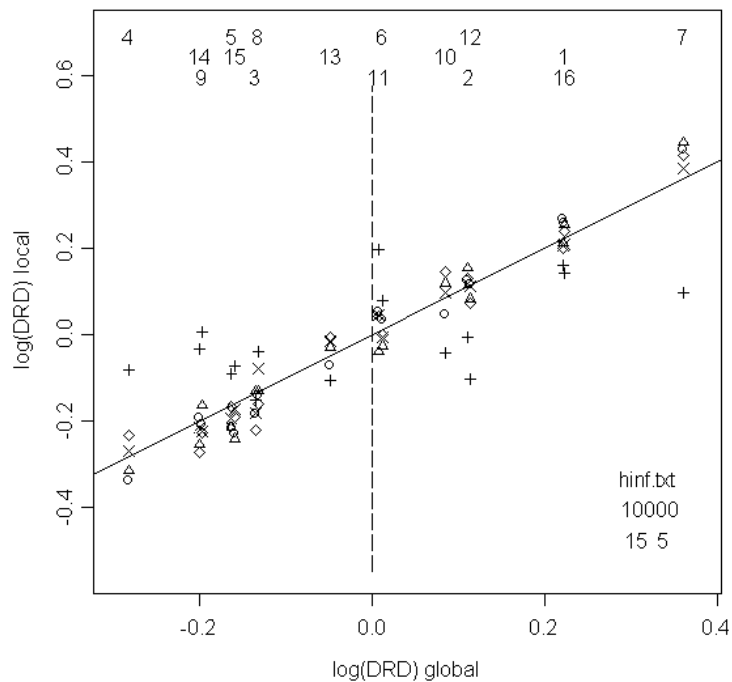


Figure 6. A log-log plot of local versus global DRD components for five contiguous 10 kilobase (kb) windows on the *H. influenzae* sequence beginning 150 kb from sequence start.

The behavior illustrated by Figure 6 is widely repeated throughout the *H. influenzae* sequence. The 16 components of the local DRD are readily classified as being rotated clockwise or counter-clockwise with respect to the line of unit slope through the origin of the log-log plot. The counter-clockwise number and the clockwise number thus sum to 16. (Points that lie exactly on the line are classified as clockwise.) Equation (5) implies that the counter-clockwise number k will increase (or decrease) from 8 when β is greater (less) than 1. To test this claim, the slope β of the regression line for each segment is computed by the method of least squares. Figure 7 shows this estimated β versus k for all 5 kb segments of the sequence. The plane is divided into four quadrants by the lines $\beta = 1$ and $k = 8$. Concentration of the plotted points in the lower left and upper right quadrants seems to substantiate the claim. If L denotes the number of points so concentrated, and M is the total number of segments in the sequence, then "odds ratios" $OR = L/(M - L)$ larger than 1 support the claim. (*H. influenzae* has $OR = 8.69$ as noted near the lower right corner of the plot.) Let the null hypothesis be that k and the estimated β are statistically independent. Counting up the points by quadrant gives a 2-by-2 contingency table with respect to which the null hypothesis can be subjected to a chi-squared test. For the *H. influenzae* sequence, the null hypothesis is rejected with P-value less than 10^{-16} .

The analysis and heuristic test procedure just described were applied to all 26 sequences to yield results in Table 3. The odds ratio exceeds 2.0 in every case and exceeds 3.0 with six exceptions. The base ten logarithm of the P-value from the chi-squared test of independence is listed as -16 whenever it is actually smaller. All test results are significant at the 10^{-5} level. The four largest P-values coincide with four of the six smallest odds ratios (for *B. burgdorferi*, *Mycoplasma genitalium*, and the two *P. falciparum* chromosomes). The fifth data column in Table 3 lists standard deviations of the estimated β , denoted $\sigma(\beta)$, which range from 0.123 (*Helicobacter pylori*) to 0.369 (*A. thaliana*). The mean of all the local estimates of β is close to 1 in every case, ranging from 0.985 (*Salmonella enterica*) to 1.065 (human chromosome XXII).

5.3. Linear relation. Equations (5) and (2) suggest an increasing, linear relation between SMI and β . The increasing relation is confirmed in *H. influenzae* by Figure 8, a plot of estimated β versus $2000I$ in 5 kb segments, but the linearity is imperfect as the points appear to form an arc through the regression line. The residual standard error 0.0702 is less than half of $\sigma(\beta)$ and the simple correlation coefficient, computed with a 5% trim to suppress the influence of outliers, is 0.951 as printed on the graph. (Better linearity is actually obtained from the regression on \sqrt{I} as the residual standard error drops to 0.0603 and the correlation coefficient rises to 0.964.) Correlation coefficients (with 5% trim) for all 26 sequences are listed in the right-most column of Table 3. Since the median of these values is 0.886, the correlation in *H. influenzae* is comparatively strong. A correlation coefficient below 0.700 corresponds to a multiple R-squared regression diagnostic less than 0.490 and hence to a general lack of fit. Six sequences give $\text{cor}(I, \beta) \leq 0.701$. Not surprisingly, these are the same six cases which produced odds ratios below 3.0.

Another view of the association between SMI and β can be obtained by plotting the cumulative sums of $I - \text{mean}(I)$ and $\beta - \text{mean}(\beta)$ versus position in

the sequence on the same graph sheet. This view of the *H. influenzae* sequence is shown in Figure 9. Regions where the measurement is suppressed relative to the global mean trend downward trends and regions of enhancement trend upward. As

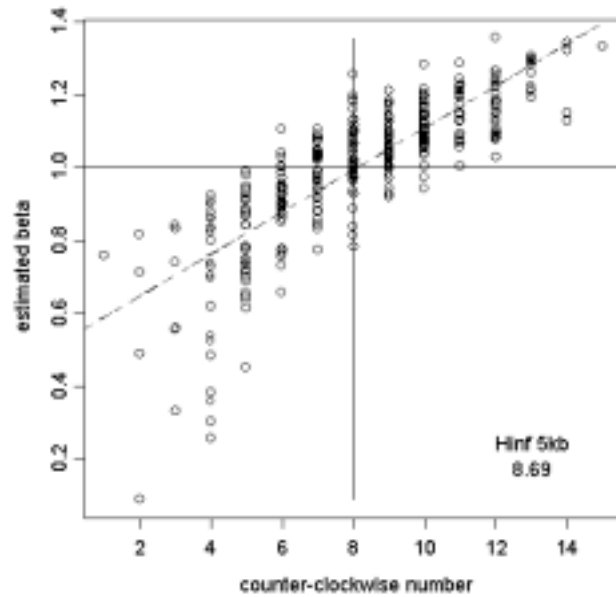


Figure 7. Estimated β versus counter-clockwise number in 5 kb segments of the *H. influenzae* sequence. A counter-clockwise rotation of the log-log plot of local versus global DRD is implied by $\beta > 0$ (and a clockwise rotation when negative).

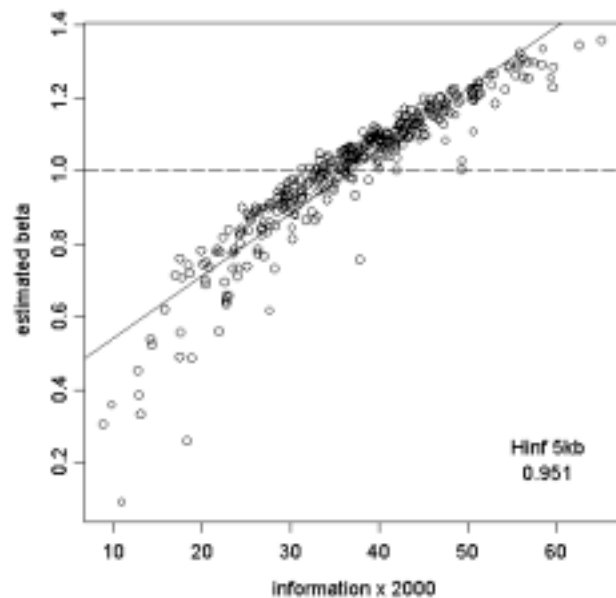


Figure 8. Estimated β versus SMI in 5 kb segments of the *H. influenzae* sequence. An increasing, linear relation between is implied by the adjusted model.

with the analysis of compositional heterogeneity in genomes, this kind of integral representation can illuminate global patterns of organization, where simple time series plots only highlight the apparent randomness.

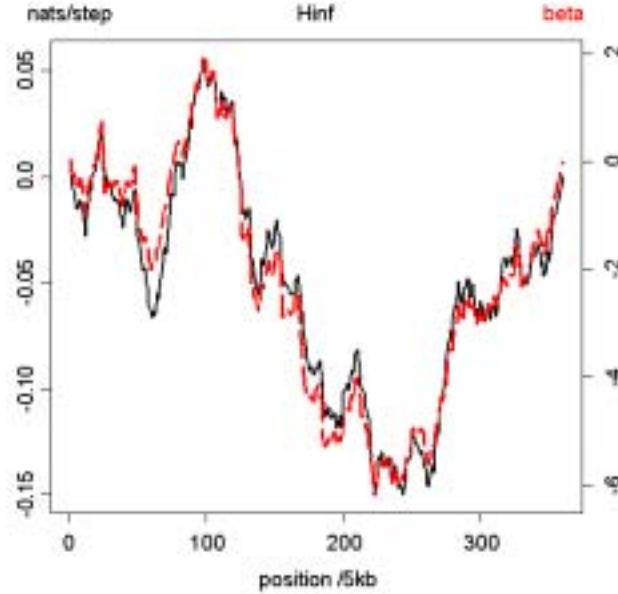


Figure 9. Cumulative local fluctuation versus position in 5 kilobase windows on the *H. influenzae* sequence. Information fluctuations (solid, black) are referred to the left scale. Fluctuations in the estimated beta (broken, red) are referred to the right scale.

5.4. Analysis of excess variation. Does the conditional Poisson model, after adjustment by equations (4) and (5), explain the excess variation and the wide dispersion of the SMI in scatter plots like Figure 5? The preceding analysis has shown that local deviations of the estimated β are mainly strongly correlated with DRD fluctuations but large excursions in these estimates (from the global average $\beta = 1$) may be associated with a lack of linearity. When this occurs the estimates may be misleading. Interpolated values of the local DRD components, obtained from (4), are compared to collocated observations by computing the variation $(1/16)\sum\sum|r_{ij} - r_{ij}(\beta)|$ as a fraction of δ . This fraction, expressed as a percentage, ranges from 3.1% (*M. genitalium*) to 13.6% (*B. subtilis*) with a median of 8.5%. These results for 5 kb segments are fairly insensitive to segment size with the median increasing to 8.8% for 25 kb and decreasing to 7.1% for 1 kb segments.

In the absence of a reliable estimate of β , neither equation (4) nor (5) can be computed to a specified accuracy, but the global average frequencies can be substituted for the local composition to obtain

$$(6) \quad I_g(\beta) = \beta I_g(1) - \log(\sum_i \sum_j \rho_{ij}^\beta g_i g_j)$$

for the SMI and

$$(7) \quad \delta(\beta) = (1/16) \sum_i \sum_j \left| \frac{\rho_{ij}^\beta}{D_g(\beta)} - \rho_{ij} \right|$$

for the variation where $D_g(\beta)$ is the double sum in (6). As the unknown parameter β is swept through a range from 0 to 8 (dimensionless units), the points $(2000I_g, \delta_g)$ form a curve that is convex (concave) on the left (right) side of the minimum at $\beta = 1$. In Figure 5, for *H. influenzae*, the fitted curve neatly supports the plotted points. As the parameter is swept through the range 0 to -8, the last two equations trace a concave (increasing) curve that arcs over the plotted points. This upper branch (only a small piece of which is visible in the upper left of Figure 5) indicates how large the variation can become when local DRD is systematically inverted in the sense of being less than 1 when the global DRD exceeds 1 (or *vice versa*). The lower branch, obtained with positive exponents, will be called the *supporting contour*.

How much excess variation is absorbed by the supporting contour? This question is answered by evaluating the height of every plotted point above the contour. The height is smaller than the variation by $\delta - \delta(\beta)$ and the supporting contour thus absorbs $100\delta(\beta)/\delta$ per cent of the variation. This absorption percentage was computed for each segment of each sequence using an algorithm that discretizes the abscissa $2000I$ in increments of 0.04. The average absorption percentages in 5kb segments are listed for each sequence in Table 2. They range from 18% (*B. burgdorferi*) to 55% (*P. falciparum* chromosome III) with a median of 31%.

If the adjusted model were a complete explanation of the phenomena then the excess variation ought to be accounted for by local modulation of the exponent. The excess variation is the average of $\delta - \delta^*$ in segments with δ^* predicted by the unadjusted model. After adjustment we expect the average values of $\delta - \delta(\beta)$ and δ^* to be about the same. Thus the observed mean value (a) of 1000δ is adjusted by (c) the absorption percentage and decremented by (b) the mean $1000\delta^*$ to obtain the residuals in the last column of Table 2. In Figure 2, where the observations were plotted against (b) $1000\delta^*$ as solid triangles, the adjusted observations are included as open circles. While the unadjusted observation always exceeded prediction by at least 35%, the adjusted observation never exceeds prediction by more than that amount. The four largest residuals coincide with the four smallest absorption percentages (*B. burgdorferi*, *Methanococcus jannaschii*, and the *Mycoplasma* sequences).

REFERENCES

- Agresti A, 1990. *Categorical Data Analysis*, New York: Wiley.
 Avery PJ and DA Henderson, 1999. Fitting Markov chain models to discrete state series such as DNA sequences. *Applied Statistics* 48: 53-61.

- Campbell A, Mrázek J, and S Karlin, 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 96: 9184-9189.
- Christensen OF and R Waagepetersen, 2002. Bayesian prediction of spatial count data using generalized linear mixed models, *Biometrics* 58: 280-286.
- Grosse I, Herzel H, Buldyrev SV, and HE Stanley, 2000. Species independence of coding and noncoding DNA. *Phys. Rev. E* 61: 5624-5629.
- Jernigan RW and RH Baran, 2002. Pervasive properties of the genomic signature, *BMC Genomics* 3: 23.
- Karlin S and SF Altschul, 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87: 2264-2268.
- Karlin S and V Brendel, 1993. Patchiness and correlations in DNA sequences. *Science* 259: 667-679.
- Karlin S, Landunga I, and BE Blaisdell, 1994. Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA* 91: 12837-12841.
- Karlin S and J Mrázek, 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 94: 10227-10232.
- Pevzner PA, 1992. Nucleotide sequences versus Markov models. *Computers Chem.* 16: 103-106.
- Robin S and J-J Daudin, 2001. Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.* 4: 895-905.
- Román-Roldán R, Bernaola-Galván P, and JL Oliver, 1996. Application of information theory to DNA sequence analysis: A review. *Pattern Recognition* 29:1187-1194.
- Stanley HE, 2000. Exotic statistical physics: Applications to biology, medicine, and economics. *Physica A* 285: 1-17.
- Yuan A and B Clarke, 1999. An information criterion for likelihood selection. *IEEE Trans. Information Theory* 45: 562-571.