Divergence and Shannon Information in Genomes

Hong-Da Chen,[1] Chang-Heng Chang,[1] Li-Ching Hsieh,[4] and Hoong-Chien Lee[1,2,3]

[1]*Department of Physics, National Central University, Chungli, Taiwan 320, Republic of China*
[2]*Department of Life Science, National Central University, Chungli, Taiwan 320, Republic of China*
[3]*Center for Complex Systems, National Central University, Chungli, Taiwan 320, Republic of China*
[4]*Institute of Information Science and Genomics Research Center, Academia Sinica, Taipei, Taiwan 115, Republic of China*
(Received 23 September 2004; published 5 May 2005)

Shannon information (SI) and its special case, divergence, are defined for a DNA sequence in terms of probabilities of chemical words in the sequence and are computed for a set of complete genomes highly diverse in length and composition. We find the following: SI (but not divergence) is inversely proportional to sequence length for a random sequence but is length independent for genomes; the genomic SI is always greater and, for shorter words and longer sequences, hundreds to thousands times greater than the SI in a random sequence whose length and composition match those of the genome; genomic SIs appear to have word-length dependent universal values. The universality is inferred to be an evolution footprint of a universal mode for genome growth.

Shannon entropy [1] has been used in almost every field concerned with information, including the analysis of DNA and protein sequences [2–5]. In the field of comparative genomics, however, it seems not to have found any systematic application. The high heterogeneity of complete genomes in length, base composition, and percentage of coding regions may make comparison based on Shannon entropy problematic. Here we show that, by simple and appropriate definition of a quantity we call *Shannon information* (SI), the difficulties associated with these issues are surmounted. The SIs are applied to characterize the distribution of the occurrence frequency of k-mers, or words of a k chemical letter, in complete genomes. We present a simple relation between the SI and the relative spectral width of a distribution, a relation that furnishes an intuitive understanding between SI and information in a sequence. We show that, in spite of their high heterogeneity, the SIs in complete genomes can be represented by a set of genome independent universal lengths.

Divergence and Shannon information.—Consider a set of occurrence frequencies, $\mathcal{F} = \{f_i | \sum_{i=1}^{\tau} f_i = L\} \equiv \{f_i | \tau, L\}$, for τ types of events, where f_i/L is the probability of event i. The Shannon entropy, or uncertainty [1], for $\mathcal{F}$ is $H(\mathcal{F}) = -\sum_i (f_i/L) \ln(f_i/L)$. The quantity attains its maximum value $H_{\max} = \ln\tau$ when all f_i are equal to the mean frequency $\bar{f} = \tau^{-1}L$. In H, Shannon was concerned with the fidelity of messages as they are transported through communication devices. Here we are interested in the information in $\mathcal{F}$ itself. There is a general notion that information in a system increases with a decrease in uncertainty; hence we identify the quantity

$$D(\mathcal{F}) \equiv \ln\tau - H(\mathcal{F}) = L^{-1} \sum_i f_i \ln(f_i/\bar{f}), \quad (1)$$

called *divergence* by Gatlin [2], as the zeroth order SI in $\mathcal{F}$. Shannon called the ratio $D(\mathcal{F})/H_{\max}$ redundancy.

Suppose the set $\mathcal{F} = \{\mathcal{F}_m | m = 1, 2, \cdots\}$ is composed of subsets, $\mathcal{F}_m = \{f_i | \tau_m, L_m\}$, each having its own distinct type of events τ_m, total number of events L_m, and mean frequency $\bar{f}_m = L_m/\tau_m$, where $\sum_m \tau_m = \tau$ and $\sum_m L_m = L$. The divergence of each subset is $D(\mathcal{F}_m) = \sum_i' f_i / L_m \ln(f_i/\bar{f}_m)$, where the summation is restricted to those f_i's in the subset $\mathcal{F}_m$. We define the SI carried by $\mathcal{F}$ to be the weighed average of the divergences in the subsets:

$$R(\mathcal{F}) \equiv \sum_m (L_m/L) D(\mathcal{F}_m) = L^{-1} \sum_{m,i}' f_i \ln(f_i/\bar{f}_m). \quad (2)$$

Frequency distribution in a DNA sequence.—We view a single strand of DNA as linear text written in the four chemical letters, A, C, G, and T, representing the four kinds of nucleotides. Empirically, genomes are invariably within a few percent of being compositionally self-complementary, and, for the present study, it suffices to characterize the base composition of a genome by a single number, p, the combined probability of (A + T). From now on, the term *profile* of a sequence will refer to the p value and the length L of the sequence. We will use the SI in random sequences as benchmarks for the SI in genomes. A random sequence having the profile of a genome is said to be a *random match* for the genome.

For a DNA sequence of length L, we denote by $\mathcal{F}^{\{k\}}$ the set $\{f_i | \tau = 4^k, L\}_k$, where f_i is the occurrence frequency of the ith overlapping a k-letter word, or k-mer [6]. For $\mathcal{F}^{\{k\}}$ the number of event types is $\tau = 4^k$ and the mean frequency is $\bar{f} = 4^{-k}L$. Given $\mathcal{F}^{\{k\}}$, we can construct a k *spectrum* where n_f, the number of k-mers occurring with frequency f, is given as a function of f. For simplicity, we also call $\mathcal{F}^{\{k\}}$ a k spectrum.

The $\mathcal{F}^{\{k\}}$ of a genomes is always composed of $k + 1$ subsets $\mathcal{F}_m^{\{k\}}$, called m sets, $m = 1$ to k, where $\mathcal{F}_m^{\{k\}}$ is the subset of k-mers with m (A + T)'s. When a genome does not have $p \approx 0.5$, and most genomes are of this type, then
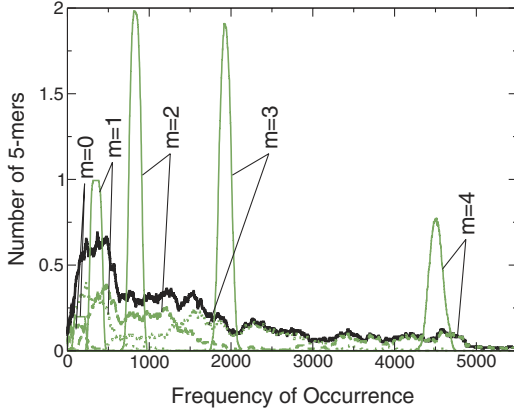
FIG. 1 (color online). The 5-spectra of the genome of
C. acetobutylicum (black lines) ($p \approx 0.7$) and that of its random
match (green or gray lines) normalized to that of a 2 Mb
sequence. The random-match spectrum is composed of sharply
peaking subspectra $\mathcal{F}_m^{\{k\}}$, $m = 0$ to 4. The center of the $m = 5$
subspectrum is off scale at 10 500. Ordinates n_f are not integers
because (for better viewing only) large fluctuations in the spectra
are smoothed out by averaging over a range of 21 frequencies.

the mean frequencies $\bar{f}_m$ of the subsets $\mathcal{F}_m^{\{k\}}$ are dis-
tinct: $\bar{f}_m(p) = \bar{f} 2^k p^m (1-p)^{k-m}$. Figure 1 [7] shows the
5-spectra of a $p = 0.691$ genome, *Clostridium acetobuty-
licum*, and its random match. The spectrum from the
random match [sharp peaks in gray (or green)] is composed
of five nonoverlapping subspectra (a sixth one is out of
range), one for each $\mathcal{F}_m^{\{5\}}$. The corresponding subspectra
(dash and dotted gray curves) underlying the entire ge-
nomic spectrum (black curve) are also shown; they are seen
to have overlapping widths.

The types and number of k-mers in $\mathcal{F}_m^{\{k\}}$ are $\tau_m = 2^k(k, m)$ and $L_m = \tau_m \bar{f}_m$, respectively, where (k, m) is
the binomial coefficient. As in Eq. (2), we define the SI
in $\mathcal{F}^{\{k\}}$ as the weighed mean of the divergences in $\mathcal{F}_m^{\{k\}}$:
$R(\mathcal{F}^{\{k\}}) = \sum_{m,i}' f_i / L \ln(f_i / \bar{f}_m)$. When p approaches 0.5,
all $\bar{f}_m$ tend to $\bar{f}$ and $R(\mathcal{F}^{\{k\}})$ approaches $D(\mathcal{F}^{\{k\}})$. Note that
$R(\mathcal{F}^{\{k\}})$ and $D(\mathcal{F}^{\{k\}})$ are invariant under the replacement
$p \to 1 - p$.

Relative spectral width.—In Eq. (1), by expanding f_i
around $\bar{f}$ in the logarithm, we obtain for a unimodal $\mathcal{F}$ the
series

$$D(\mathcal{F}) = \sum_{n=2} \frac{(-1)^n}{n(n-1)} \left\langle \left(\frac{f_i - \bar{f}}{\bar{f}} \right)^n \right\rangle \approx \frac{\Delta^2}{2\bar{f}^2}, \quad (3)$$

where Δ is the standard deviation in $\mathcal{F}$, or $D(\mathcal{F}) \approx \sigma^2/2$,
where σ is the *relative spectral width* (RSW) of $\mathcal{F}$. This
relation is useful in at least two respects. First, it gives one
a heuristic understanding why the divergence of and infor-
mation in a (unimodal) spectrum are connected. When the
spectrum is narrow, there is little information because all
the k-mers occur with almost equal frequency. Because of
the generally substantiated notion that highly overrepre-
sented words in a genome are more likely to have biologi-
cal meaning than words represented with an average
frequency [8–11], the obverse is also true; namely, there
is more information when the spectrum is broad. Second,
Eq. (3) gives an accurate estimate of the divergence in the
$\mathcal{F}_m^{\{k\}}$ (and $\mathcal{F}^{\{k\}}$) of a random sequence. Provided $\bar{f}_m$ is
much greater than 1, $\mathcal{F}_m^{\{k\}}$ is proportional to a Poisson dis-
tribution with mean $\bar{f}_m$ [12,13], so that $\Delta_{\text{ran}}^2 \approx b_k \bar{f}_m$ and
$\sigma_{\text{ran}}^2 \approx b_k / \bar{f}_m$. Hence $D(\mathcal{F}_m^{\{k\}}) \approx b_k / 2\bar{f}_m = b_k \tau_m / 2L_m$
for a random sequence. Then, from Eq. (2) we have a re-
sult for the SI in the $\mathcal{F}^{\{k\}}$ of any random match, indepen-
dent of p:

$$R(\mathcal{F}^{\{k\}}) \approx b_k \tau / 2L \quad \text{(Random sequence).} \quad (4)$$

The combinatorial factor b_k should be $1 - \tau^{-1}$ for a true
random sequence; a semiempirical value $1 - 1/2^{k-1}$ is
used here because a random match is not fully random:
it is made to be approximately compositionally self-
complementary (as most genomes are) and its p value is
fixed. That the SI in a random sequence diminishes as $1/L$
with increasing L is connected to what is known as the
large-system rule: the square of the RSW of a distribution
associated with a random system is inversely proportional
to the size of the system. We will see that the SI in genomes
does not follow this rule.

TABLE I. Shannon entropy H and divergence D in units of ln2 in the k spectra of the genome
sequence of *E. coli 0157:H7* and its random match.

	Random match			*E. coli 0157:H7*		
k	H	D	$b_k 4^k / 2L$	H	D	$\sigma^2/2$
2	4.0000	5.12×10^{-7}	1.06×10^{-6}	3.986	1.39×10^{-2}	1.39×10^{-2}
3	6.0000	6.76×10^{-6}	6.26×10^{-6}	5.953	4.71×10^{-2}	4.41×10^{-2}
4	8.0000	3.03×10^{-5}	2.92×10^{-5}	7.908	9.16×10^{-2}	8.65×10^{-2}
5	9.9999	1.28×10^{-4}	1.25×10^{-4}	9.854	1.46×10^{-1}	1.42×10^{-1}
6	11.999	5.23×10^{-4}	5.18×10^{-4}	11.79	2.07×10^{-1}	2.11×10^{-1}
7	13.998	2.12×10^{-3}	2.10×10^{-3}	13.72	2.73×10^{-1}	2.90×10^{-1}
8	15.991	8.56×10^{-3}	8.48×10^{-3}	15.65	3.49×10^{-1}	3.87×10^{-1}
9	17.965	3.45×10^{-2}	3.42×10^{-2}	17.54	4.54×10^{-1}	5.22×10^{-1}
10	19.858	1.42×10^{-1}	1.37×10^{-1}	19.34	6.58×10^{-1}	7.70×10^{-1}

TABLE II. Divergence [Eq. (1)] in the m sets of the six spectra of three genomes and their random matches.

m	$\bar{f}_m$	L_m (kb)	$b_6/(2\bar{f}_m)$	D_{ran}	D_{gen}
Pyrobaculum aerophilum, $p = 0.486$, $L = 2.22$ Mb					
0	480	30.7	1.01×10^{-3}	8.54×10^{-4}	4.32×10^{-2}
1	500	192	9.69×10^{-4}	1.06×10^{-3}	6.88×10^{-2}
2	520	499	9.31×10^{-4}	9.58×10^{-4}	8.71×10^{-2}
3	541	692	8.96×10^{-4}	8.12×10^{-4}	1.33×10^{-1}
4	563	540	8.93×10^{-4}	8.80×10^{-4}	1.80×10^{-1}
5	586	225	8.60×10^{-4}	9.37×10^{-4}	1.58×10^{-1}
6	610	39.0	7.94×10^{-4}	7.65×10^{-4}	7.74×10^{-2}
Chlamydia muridarum, $p = 0.597$, $L = 1.07$ Mb					
0	68.7	4.40	7.05×10^{-3}	4.87×10^{-3}	1.11×10^{-1}
1	103	39.5	4.70×10^{-3}	4.47×10^{-3}	1.16×10^{-1}
2	155	148	3.12×10^{-3}	3.18×10^{-3}	1.15×10^{-1}
3	232	297	2.08×10^{-3}	2.15×10^{-3}	1.07×10^{-1}
4	348	334	1.39×10^{-3}	1.49×10^{-3}	1.02×10^{-1}
5	521	200	9.29×10^{-4}	1.12×10^{-3}	1.20×10^{-1}
6	782	50.0	6.14×10^{-4}	7.36×10^{-4}	1.52×10^{-1}
Streptomyces avermitilis, $p = 0.293$, $L = 9.03$ Mb					
0	89.0	5.70	5.44×10^{-3}	5.99×10^{-3}	1.73×10^{-1}
1	214	82.2	2.26×10^{-3}	2.27×10^{-3}	1.80×10^{-1}
2	519	498	9.33×10^{-4}	1.04×10^{-3}	4.21×10^{-1}
3	1252	1602	3.87×10^{-4}	3.78×10^{-4}	2.59×10^{-1}
4	3024	2903	1.60×10^{-4}	1.66×10^{-4}	1.33×10^{-1}
5	7300	2803	6.63×10^{-5}	6.14×10^{-5}	6.22×10^{-2}
6	17 622	1127	2.75×10^{-5}	2.87×10^{-5}	8.87×10^{-2}

Shannon information in a $p \approx 0.5$ genome.—The genome of *E. coli 0157:H7* is 5.53 Mb long and has $p = 0.496$. We therefore use Eq. (1) to compute its SI. The Shannon entropy and divergence (in units of ln2) in the k spectra, $k = 2$ to 10, of the genome and its random match are given in Table I. We notice the following: (i) For both sequences the Shannon entropy (columns 2 and 5) is, in every case, close to its maximum value, $2k$. (ii) For the random match the value computed using Eq. (1) (column 3) is in excellent agreement with the expected value (column 4). (iii) For the smallest k's, the divergence is a minuscule signal buried in a huge Shannon entropy background. (iv) This tiny signal, when isolated, cleanly sets apart a genome from its random match: the genomic divergence is greater than its random-match counterpart in all cases and, for the smaller k's, many thousand times greater. (v) Equation (3) is verified (last column).

Difference between divergence and SI.—When p is not close to 0.5, $D(\mathcal{F}^{\{k\}})$ and $F(\mathcal{F}^{\{k\}})$ differ. Consider the 5-spectra (not the subspectra) for the genome and its random match shown in Fig. 1. Because the $\bar{f}_m$'s are spread widely, their distribution plays a dominant role in determining the width of the spectra. If we ignore the widths of the individual subspectra and use Eq. (3), we get $D(\mathcal{F}^{\{k\}}) \approx D^{(0)}(k, p) = 0.5(\{2^k[p^2 + (1 - p)^2]\}^k - 1)$ for both the genomic and random spectra $[D^{(0)}(5, 0.691) = 0.488$, as

compared to the actual values of 0.575 and 0.485 for the genomic and random 5-spectra in Fig. 1]. $D(\mathcal{F}^{\{k\}})$ depends strongly on p, is independent of L, disagrees with Eq. (4), and is therefore not useful for estimating the SI of a random sequence.

The monomer divergence and SI are quantities that are completely determined by the base composition of a sequence and are independent of its randomness. Whereas $R(\mathcal{F}^{\{1\}}) = 0$ always, $D(\mathcal{F}^{\{1\}})$ depends strongly on p and vanishes only at $p = 0.5$. The genome of *Mycoplasma genitalium* has $p = 0.683$ and $D(\mathcal{F}^{\{1\}}) = 0.0686$ (compared with $D^{(0)} = 0.670$). The genome *Haemophilus influenzae* has $p = 0.618$ and $D(\mathcal{F}^{\{1\}}) = 0.0281$ (0.0278). Since these values say nothing about the randomness or order of the sequence, yet depend nontrivially on p, they convey *less* information about the sequence than the value of p alone.

The general case.—Table II gives the divergences D_{gen} and D_{ran} of the m sets in the 6-spectra of three sequences and their random matches, *P. aerophilum*, *C. muridarum*, and *S. avermitilis*. These sequences span a wide range in profile. Once again, D_{ran} (column 5) is well approximated by its expected value (column 4) and is much less than D_{gen}. D_{gen} varies but does not exhibit a clear dependence on L. The 6-mer SI, computed according to Eq. (2), for *E. coli 0157:H7* and the three genomes of Table II and their

TABLE III.   Shannon information [Eq. (2)] in the six spectra of the four genomes cited in Tables I and II and their random matches.

Genome	p	L (Mb)	$\bar{f}$	R_{ran}	$R_{\mathrm{ran}}/(b_6/2\bar{f})$	R_{gen}
E. coli	0.486	5.52	1348	3.59×10^{-4}	1.01	0.143
P. aerophilum	0.514	2.22	542	8.97×10^{-4}	0.995	0.127
C. muridarum	0.597	1.07	261	1.83×10^{-3}	1.05	0.108
S. avermitilis	0.293	9.03	2203	2.20×10^{-4}	1.03	0.144

random matches are given in Table III. Again, we have R_{ran} well approximated by $b_k \tau / 2L$, $R_{\mathrm{gen}} \gg R_{\mathrm{ran}}$, and R_{gen} independent of genome length.

A new feature is manifest in Table III: in spite of their greatly varying profiles, the four genomes have almost the same SI in their 6-spectra. Indeed, at least for word lengths up to 10 letters, we have found that SI varies little for complete genomes with vastly differing profiles. Figure 2 shows SI as a function of sequence length in six prokaryotes and six eukaryotic chromosomes [7].

Genomes have universal SI.—With the exception of several 2-mer SIs, all genomic SIs lie within a horizontal band bounded by 0.01 and 0.9. (For technical reasons too lengthy to describe here but which will be explained elsewhere, the 2-mer SI has the largest fluctuations.) For a given k, the genomic variance in SI is only about a factor

of 2. In comparison, the longest sequence—*H. sapiens* (*I*) at 228 Mb—is about 1200 times longer than the shortest—*E. cuniculi* (*I*) at 0.198 Mb. In this context, we view the genomic SI as having a genome-independent but k-dependent universal value. For each k, we express this universal value, $R_{\mathrm{uni}}(k)$, in terms of the length, L_r, called the *equivalent root sequence*, of a random sequence whose SI is equal to R_{uni}: $L_r(k) \equiv 4^k b_k / 2R_{\mathrm{uni}}(k)$. Then the genomic results in Fig. 2 are summarized by the empirical relation $\ln[L_r(k)] = \alpha k + C$, where $\alpha = 1.01 \pm 0.06$ and $C = 3.80 \pm 0.50$. L_r is surprisingly short for small k but grows rapidly with k: $L_r(k) \approx 340$ b, 15 kb, and 1.1 Mb for $k = 2$, 6, and 10, respectively.

Now, if a random root sequence of length L_r is replicated a finite number of times, then, provided $k \ll L_r$, the k-mer SI in the product sequence is the same as that in the root sequence. We therefore suggest that the observed universality of SI in complete genomes is an evolution footprint produced by a universal mode of genome growth, a mode that began when the genomes were very short—less than 340 b—and in which segmental duplication played a major role [13,14]. Work exploring the implication of this notion will be reported elsewhere.
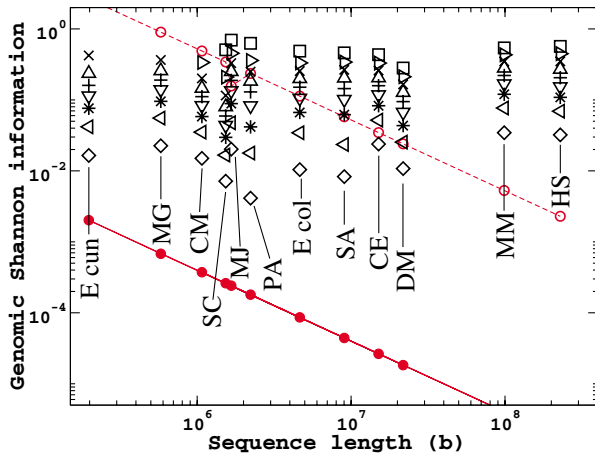
FIG. 2 (color online).   SI in the k spectra for $k = 2$ ($\diamond$), 3 ($\triangleleft$), 4 (*), 5 ($\triangledown$), 6 (+), 7 ($\triangle$), 8 ($\times$), 9 ($\triangleright$), and 10 ($\square$) of the genomes of 12 organisms: E cun, *E. cuniculi* (*Chromosome I*; length 0.198 Mb, $p = 0.533$); CM, *C. muridarum*; MG, *M. genitalium* (0.580 Mb, 0.683); SC, yeast (*S cerevisiae, IV*; 1.53 Mb, 0.621); MJ, *M. jannaschii* (1.66 Mb, 0.686); PA, *P. aerophilum*; E col, *E. coli K12* (4.64 Mb, 0.492); SA, *S. avermitilis*; CE, worm (*C. elegans, I*; 15.1 Mb, 0.643); DM, fly (*D. melanogaster, X*; 21.8 Mb, 0.574); MM, mouse (*M. musculus, IX*; 98.9 Mb, 0.562); HS, human (*H. sapiens, I*; 228 Mb, 0.582). The lines give SI expected in random sequences for $k = 10$ (open circles connected by dotted line) and 100 times the expected value for $k = 2$ (solid circles connected by solid line). The k-mer SI of a sequence is not given when sequence length is less than twice 4^k.

[1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).

[2] L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).

[3] P. Clote and R. Backofen, *Computational Molecular Biology: An Introduction* (John Wiley & Sons, New York, 2000).

[4] P. Bernaola-Galván *et al.*, Phys. Rev. Lett. **85**, 1342 (2000).

[5] Ivo Grosse *et al.*, Phys. Rev. E **65**, 041905 (2002)

[6] B. L. Hao, H. C. Lee, and S. Y. Zhang, Chaos Solitons Fractals **11**, 825 (2000).

[7] All genome sequences are downloaded from GenBank; prokaryotes from http://www.ncbi.nlm.nih.gov/genomes/Complete.html and eukaryotes from http://www.ncbi.nlm.nih.gov/genomes/static/euk_g.html.

[8] H. O. Smith *et al.*, Science **269**, 538 (1995).

[9] S. Karlin, J. Mrazek, and M. Campbell, Nucleic Acids Res. **24**, 4263 (1996).

[10] J. van Helden, B. Andre, and J. Collado-Vides, J. Mol. Biol. **281**, 827 (1998).

[11] H. J. Bussemaker, H. Li, and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 10 096 (2000).

[12] H. M. Xie and B. L. Hao, in *Proceedings of the 2003 IEEE Computational Systems Bioinformatics Conference, Stanford, CA* (IEEE Report No. CSB2003, 2003), p. 31.

[13] L. C. Hsieh *et al.*, Phys. Rev. Lett. **90**, 018101 (2003).

[14] Depository research articles by L. C. Hsieh *et al.*, Genome Biol. **4**, P7 (2003); **5**, P7 (2004).