ELSEVIER

Biophysical Chemistry 117 (2005) 87 – 95

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Universal scaling of the C–G distribution of genes

Douglas Poland*

Department of Chemistry, The Johns Hopkins University, Baltimore, MD 21218, USA

Received 22 March 2005; accepted 23 March 2005
Available online 19 May 2005

Abstract

Using our previous result that the C–G distribution in genomes is very broad, varying as a power law of the size of the block of genome considered, we examine the C–G distribution in genes themselves. We show that the widths of the C–G distributions for the genes of several simple organisms also vary as power laws. This suggests that the power law behavior gives a universal scaling whereby the distributions for the C–G content of the genes from all species are mapped onto a single function.

Keywords: Power law; Persistence exponent; Scaling relation; DNA; Fractional Brownian walk; Fractal dimension

1. Introduction

In two previous publications [1,2] we examined the C–G distribution in single chromosomes, which for most of the species examined (specifically, the bacteria) constitute the complete genome. We divided the base sequence of the chromosome up into consecutive, nonoverlapping boxes each containing m bases as illustrated for $m=50$ and $m=100$ in Fig. 1. We then simply counted the number of bases, n, that are either C or G in each box. This process then gives an empirical distribution function for the distribution of C–G bases in m-boxes. Specifically, the probability that an m-box will contain n C–G bases is given by

$$P(n) = \text{ fraction of m} - \text{boxes that contain n C} - \text{G bases} \tag{1}$$

The first two moments of the C–G distribution function are then given empirically in terms of the $P(n)$ as follows

$$\mu_1 = \sum_{n=0}^{m} nP(n) \text{ and } \mu_2 = \sum_{n=0}^{m} n^2 P(n) \tag{2}$$

Given the first two moments of the distribution, the standard deviation, a measure of the width of the distribution, is given by the relation

$$\sigma_m = \sqrt{\mu_2 - \mu_1^2} \tag{3}$$

As a basis for comparison, we consider the case of a random distribution. We begin by defining the following average fractions

$$f_{\text{at}} = \text{fraction of bases that are A or T}$$

$$f_{\text{cg}} = \text{fraction of bases that are C or G} \tag{4}$$

with

$$f_{\text{at}} + f_{\text{cg}} = 1 \tag{5}$$

For a random distribution with the average composition given in terms of the f's above, the probability that a box of m bases contains n C–G bases is given by the standard combinatorial expression

$$P(n) = \frac{m!}{(m-n)!n!} f_{\text{at}}^{m-n} f_{\text{cg}}^{n} \tag{6}$$

Using Eq. (6) in Eq. (2) gives the standard deviation, as given by Eq. (3), of this distribution (denoted as σ_m')

$$\sigma_m' = \sqrt{f_{\text{cg}} - f_{\text{cg}}^2}\, m^{1/2} \tag{7}$$

Abbreviations: Hp, *Heliobacter pylori*; D, fractal dimension.

* Tel.: +1 410 516 7441; fax: +1 410 516 8420.
E-mail address: poland@jhu.edu.

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2005.03.012

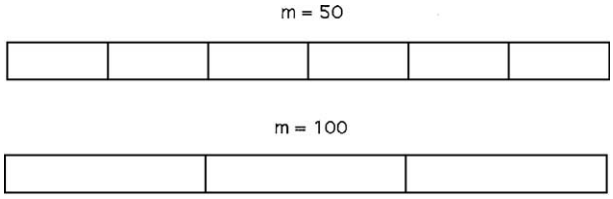

Fig. 1. Illustration of the division of a genome into consecutive boxes each containing m basepairs. Illustrated are the cases $m=50$ and $m=100$. Analysis of the C–G content of the blocks yields the box distribution for a particular genome.

What we find when we construct the empirical distribution functions for m-boxes is that the distributions obtained are very much broader than random distributions with the same net C–G content. A useful way to examine the width of the empirical distribution functions is to take them relative to the width of the random distribution given by Eq. (7). To this end, we define the function

$$\zeta(m) = \sigma_m / m^{1/2} \tag{8}$$

If the empirical distribution is random, then this function should be independent of m. To illustrate this process, we will take the genome (a single circular chromosome) of the bacteria *Heliobacter pylori* (Hp for short). The genome for this species was determined by Alm et al. [3] and can be obtained from the institute for genomic research (Tigr) on the web [4]. Hp is one of the species we have treated previously [2]. The overall length of the chromosome and fraction of C–G bases are given below

$$N = 1,643,831$$

$$f_{\mathrm{cg}} = 0.390 \tag{9}$$

In Fig. 2, we show the empirical values of $\zeta(m)$ as defined by Eq. (8) for $m=10$ to 1300 in steps of 10 bases (solid points). The dashed line is the constant value this quantity would have if the distribution were random and is given by

$$\zeta'(m) = \sigma'_m / m^{1/2} = \sqrt{f_{\mathrm{cg}} - f_{\mathrm{cg}}^2} = 0.488 \tag{10}$$

This graph vividly illustrates that the empirical distribution of C–G content in m-boxes is very much broader than that for a random distribution. Of course, we do not expect the base sequence to be random since it contains the genetic information. However, since the influence of the local information content is relatively short ranged, we would not expect much difference from a random distribution, just like the distribution of the letter "a" in a book will appear more or less random.

The graph in Fig. 2 strongly suggests a power law behavior of the form

$$\zeta(m) = A m^{\gamma} \tag{11}$$

If $\zeta(m)$ obeys such a power law, then ln $[\zeta(m)]$ plotted against ln(m) should give a straight line with slope γ. This

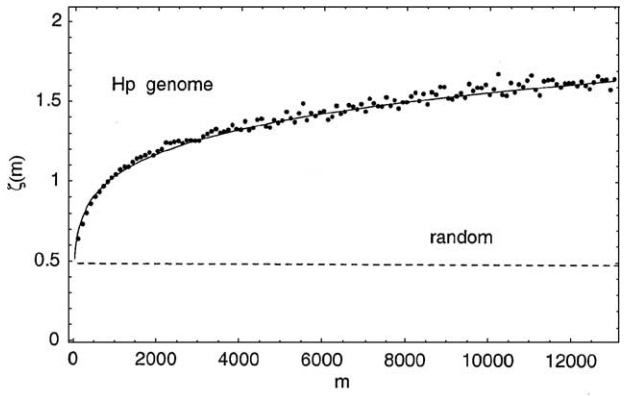

Fig. 2. The width function $\zeta(m)$ of Eq. (8) for m-boxes as a function of m for the genome of *Heliobacter pylori.* The dashed line is the constant value of ζ for a random distribution as given by Eq. (10).

plot is shown in the upper graph in Fig. 3. The straight line represents the best linear fit of the data giving the parameters for this genome

$$A = 0.303$$

$$\gamma = 0.178 \tag{12}$$

A further test is given by forming the function

$$R(m) = \zeta(m) / m^{\gamma} \tag{13}$$

If the data in Fig. 2 follow the form in Eq. (11), then this function should be a constant independent of m. This

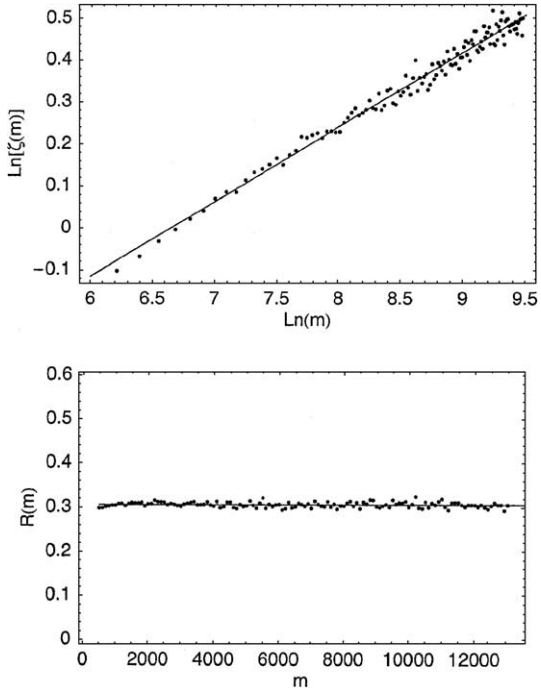

Fig. 3. Upper graph—the data of Fig. 2 plotted as ln$[\zeta(m)]$ versus ln$[m]$. The straight line is the best linear fit of the data. Lower graph—the data of Fig. 2 plotted as the function $R(m)$ of Eq. (13) as a function of m. The straight line is the constant given by Eq. (14).

function is shown in the lower graph in Fig. 3 and one sees that indeed $R(m)$ is virtually independent of m, the only variation being random fluctuations. The solid line in the graph is a plot of the constant $A=0.303$. The average value of $R(m)$ and the root-mean-square standard deviation are

$$\langle R\rangle = 0.3065 \pm 0.0058 \tag{14}$$

We have examined the m-box distribution for many species and find similar results with a characteristic exponent γ and constant A for each species [1,2]. The power law form for the C–G distributions as given by Eq. (11) implies that boxes of bases with a given C–G content tend to be followed by boxes with similar C–G content. This implies a persistence of C–G content as one goes from box to box and as a result we have named the exponent γ the persistence exponent. We have shown [1] that in order to obtain the power law behavior shown in Eq. (11) one must have correlations between m-boxes of all sizes. That is, one cannot explain the broadness of the distribution for $m=100$ boxes in terms of correlations between the C–G content for $m=50$ boxes.

We have also shown that the persistence exponent is related to the fractional random walk introduced by Mandelbrot [5]. This is also equivalent to a random walk described by the fractal dimension

$$D = 3/2 - \gamma \tag{15}$$

2. Gene C–G distributions

The remarkable correlation properties described in the previous section are for consecutive nonoverlapping m-boxes as illustrated in Fig. 1. In this paper, we will explore whether or not a similar power law distribution holds for the C–G content in genes. To do this, we will continue to use the genome of *Heliobacter pylori* as an example.

The use of m-boxes made the analysis of the C–G distribution in a given genome relatively simple in that all of the boxes were of the same size for a given value of m. For the case of the gene C–G distribution, one has the complication that the genes come in all sizes and this makes comparison more difficult. For Hp, there are NG genes listed in Tigr [4] where

$$\mathrm{NG} = 1489 \tag{16}$$

The average gene length for this species is 1104 bases long. In general, there is some spacing between consecutive genes and on occasion some overlap of genes. Thus the sum of the base content of the genes is not quite the same as that of the complete genome. We take L as the general length of a gene in bases and $\mathrm{Ng}(L)$ as the number of genes having this length. A histogram for the gene length distribution in Hp is shown in Fig. 4 where the size of the boxes used in the histogram is 200 bases; $\mathrm{Ng}(L)$ thus gives the number of genes in length windows 200 bases wide. One notes that the distribution of gene lengths given in Fig. 4 has a fairly long tail.

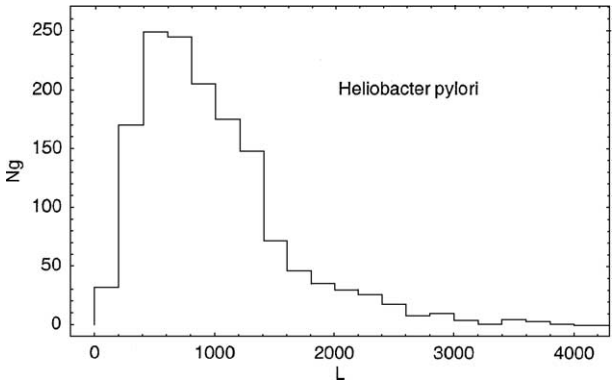

Fig. 4. Histogram for the length distribution for the genes of *Heliobacter pylori* plotting the number of genes, Ng, falling in windows of the length of the genes measured in bases; the size of the L windows is 200 bases.

One sees in Fig. 4 that most of the genes have lengths that lie in the range $L=200$ to 2000 bases. We can use the m-box distributions obtained in the previous section to get an idea of the distribution functions we expect for gene composition. We combine Eqs. (8) and (11) to give the width of the power law distribution

$$\sigma_m^* = Am^{1/2+\gamma} \tag{17}$$

This is to be contrasted with the width for the random distribution that is given by Eq. (7). These two quantities are compared in Fig. 5 for $m=10$ to 2000, the range of lengths appropriate to the Hp gene distribution. It is useful also useful to examine the actual m-box distribution functions for several values of m to compare the actual and random distributions. Assuming a Gaussian form, the distributions are given by the standard function

$$P(n) = \frac{1}{\sqrt{2\pi}\sigma_m}\exp\left[-\left(n-\overline{n}_m\right)^2/2\sigma_m^2\right] \tag{18}$$

with

$$\overline{n}_m = f_{\mathrm{cg}}m \tag{19}$$

where f_{cg} for Hp is given by Eq. (9). For the actual distribution, one uses σ_m* given by Eq. (17) while for the case of the random distribution one uses σ_m' given by Eq. (7). The random and actual distributions given by Eq. (18), using the σ_m values given respectively by Eqs. (7) and (17), for $m=25$ and $m=1000$ are shown in Fig. 6. For the case of $m=25$, one sees that there is not much difference between the random and actual distribution for Hp. Thus for small values of m, the distribution appears random. However, for the case of $m=1000$, one sees that there is an enormous difference between the much sharper random distribution and the actual distribution for the Hp genome. Thus for genes with an average size of 1104 bases, we expect that the

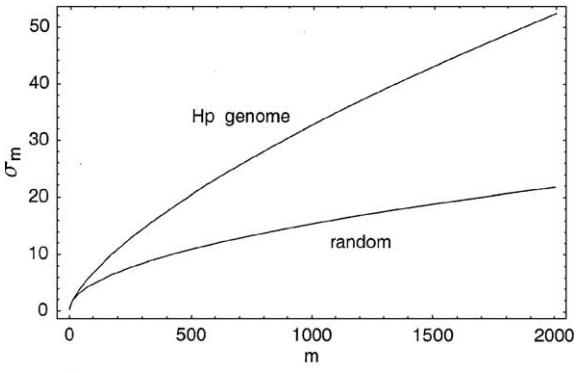

Fig. 5. Plots of the width function σ_m as a function of m. The curve marked Hp genome gives the empirical power law σ_m^* of Eq. (17) obtained from the data of Fig. 2 for *Heliobacter pylori.* The curve marked random is σ_m' of Eq. (7) for a random distribution.

C–G distribution will be very different from a random distribution. Our next task then is to construct the actual C–G distribution function for the Hp genes.

We begin our examination of the gene distribution in Hp with a survey of the gene lengths in this species. Fig. 7 shows a scatter plot of the pairs of numbers $\{L_i, N_i\}$ where L_i is the length of gene-i and N_i is the number of C–G bases in gene-i. Fig. 7 shows all of these points representing the genes of Hp except for six where L is greater than 4000. The fraction of C–G bases in gene-i is given by

$$f_i = N_i/L_i \tag{20}$$

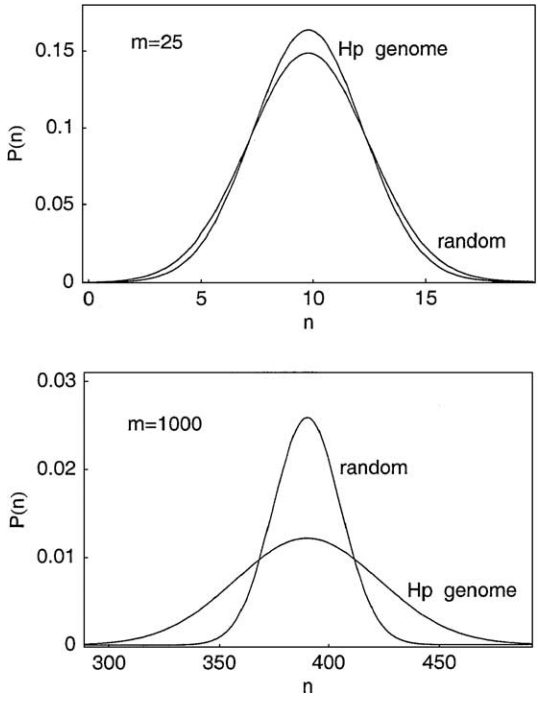

Fig. 6. Comparison of the m-box distribution given by Eq. (18) for the power law width function σ_m* of Eq. (17) and the random distribution width function σ_m' of Eq. (7). The upper graph compares the two forms of the distribution for $m=25$ while the lower graph compares the same two forms for $m=1000$.

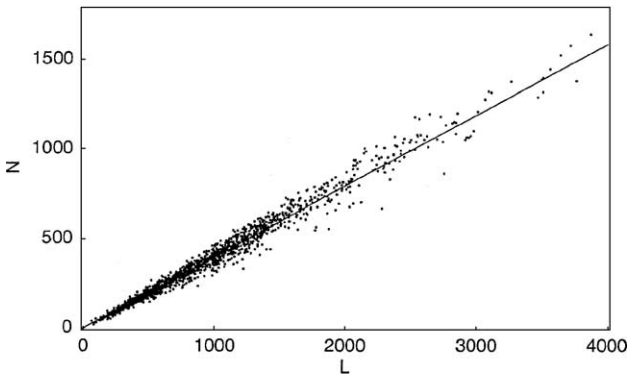

Fig. 7. Scatter plot of the sets of data $\{L_i, N_i\}$ for the genes of *Heliobacter pylori* where L_i is the length of gene-i measured in number of bases and N_i is the C–G content of that gene. The solid line is a plot of Eq. (23) giving the average N as a function of L.

while the average value of this quantity is

$$\bar{f} = \frac{1}{\mathrm{NG}} \sum_{i=1}^{\mathrm{NG}} f_i \tag{21}$$

where NG is the total number of genes given in Eq. (16). For the Hp genes, we have

$$\bar{f} = 0.396 \tag{22}$$

We note that this does not have the same value as f_{cg} given in Eq. (9) which is the overall chromosome average, not just the average for the genes.

The solid line given in Fig. 7 is the average value of N as a function of L (treating these here as continuous variables) as given by

$$\bar{N} = \bar{f}L. \tag{23}$$

where $\bar{f}$ is given by Eq. (22). We next define the difference between the actual value of N_i and the average value given by Eq. (23)

$$\Delta N_i = N_i - \bar{f}L_i \tag{24}$$

These points are plotted in Fig. 8 using the data given in Fig. 7 up to the value of $L=2000$.

We next consider the C–G distribution function for genes which we take as the analog of the Gaussian distribution for m-boxes given in Eq. (18). The length of a gene in number of bases is the analog of m, the box size while N, the number of C–G bases in a given gene, is the analog of n for m-boxes. Using the translation

$$(\text{boxes}) \quad \begin{matrix} m \to L \\ n \to N \end{matrix} \quad (\text{genes}) \tag{25}$$

we obtain the analog of Eq. (18)

$$P(N)\mathrm{d}N = \frac{1}{\sqrt{2\pi}\sigma_L} \exp\left[-\left(N - \bar{N}_L\right)^2 / 2\sigma_L^2 \right] \mathrm{d}N \tag{26}$$

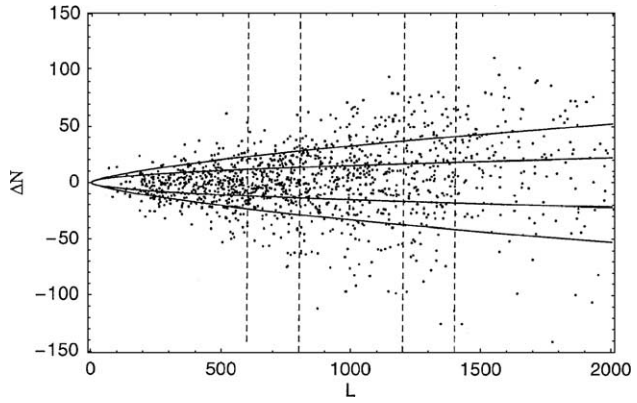

Fig. 8. A replot of the data given in Fig. 7 giving the N_i relative to the average N using the relation of Eq. (24). The inner horizontal solid curves give $\pm\sigma_L'$ for a random distribution as given by Eq. (27) while the outer horizontal solid curves give $\pm\sigma_L$* as given by Eq. (28). The vertical sets of dashed lines define bands of L-values.

where $\bar{N}_L$ is given by Eq. (23). This is the expected form of the distribution function for genes based on the behavior of the m-box distribution.

For the m-boxes, we have two expressions for σ, namely σ_m' of Eq. (7) for a random distribution and σ_m^* of Eq. (17) for the empirical power law distribution. The analogous σ's for the gene distribution are

$$\text{(random)} \quad \sigma_L' = \sqrt{\bar{f} - \bar{f}^2 L}^{1/2} \tag{27}$$

$$\text{(power law)} \quad \sigma_L^* = AL^{1/2+\gamma} \tag{28}$$

In Fig. 8, the inner solid horizontal curves give the loci of $\pm\sigma_L'$ for the random distribution given by Eq. (27) while the outer horizontal curves give the loci of $\pm\sigma_L^*$ for the power law distribution as given by Eq. (28). One clearly sees that the power law curves give a much better fit to the distribution of the data points than do the curves for the random distribution. As was the case with the m-boxes, the assumption of random behavior predicts a distribution that is much narrower than is actually found. The vertical lines in Fig. 8 represent bands of L values that we will use to illustrate the width distribution for genes. The first band encompasses the range of L=600 to 800 while the second encompasses the range of L from 1200 to 1400.

3. Universal scaling

We now consider the construction of a scaled distribution for genes. Referring to the distribution given in Eq. (26), we define the following variable

$$X = \left(N - \bar{N}_L\right)/\sigma_L \tag{29}$$

with

$$\mathrm{d}X = \mathrm{d}N/\sigma_L \tag{30}$$

We have two choices for σ_L, that of Eq. (27) for the random distribution and that of Eq. (28) for the power law distribution. For the correct form of σ_L, the resulting distribution will be independent of L and it will be the Gaussian distribution with standard deviation σ=1 as follows

$$P(X)\mathrm{d}X = \frac{1}{\sqrt{2\pi}}\exp\left[-X^2/2\right]\mathrm{d}X \tag{31}$$

We now show that the power law distribution correctly scales the gene L–N distribution. For the discrete variables L_i and N_i, we take ΔN_i of Eq. (24) and divide by the width of the power law distribution given by Eq. (28) giving

$$X_i = \left(N_i - \bar{f}L_i\right)/AL_i^{1/2+\gamma} \tag{32}$$

A plot of these data is shown in Fig. 9. The vertical dashed lines indicate the same bands of L as shown in Fig. 8. The horizontal lines indicate the loci of ±1 (which is the standard deviation of the Gaussian in Eq. (31)). One sees that the mapping of Eq. (32) now makes the width of the distribution for any L value virtually independent of the value of L. Thus the mapping of Eq. (32), using the species specific values of A and γ in Eq. (12), gives a mapping to a simple Gaussian distribution. This mapping should apply to the gene C–G distribution for the genes of any species using the appropriate values of A and γ.

The fact that the distribution of X_i values shown in Fig. 9 are virtually independent of L indicates that genes obey the scaling of Eq. (32) which in turn indicates that the C–G distributions in genes are described by the formula given in Eq. (26) with the power law width of Eq. (28). We can test the applicability of this distribution directly by examining the distribution $P(N)$ for fixed values of L. The problem in doing this is that there are a limited number of genes that have exactly the same length. That is the advantage of using m-boxes since in that case one divides up the whole genome into boxes of the same size and hence one has a large sample to work with. To partially remedy this problem, we consider bands of L values. The bands L=600–800 and

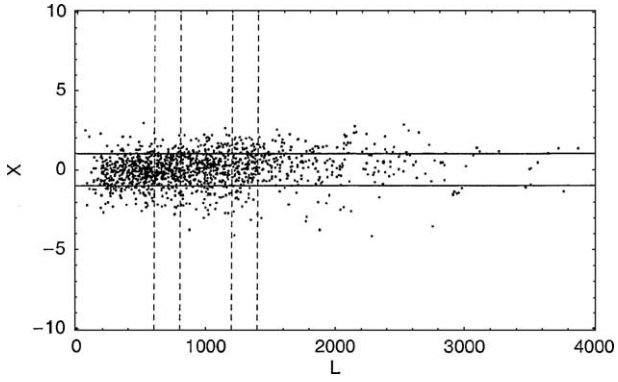

Fig. 9. A replot of the data of Fig. 7 using the function X defined in Eq. (32). The solid horizontal lines give the loci σ=±1 for the simple Gaussian distribution of Eq. (31). The vertical sets of dashed lines define bands of L-values.

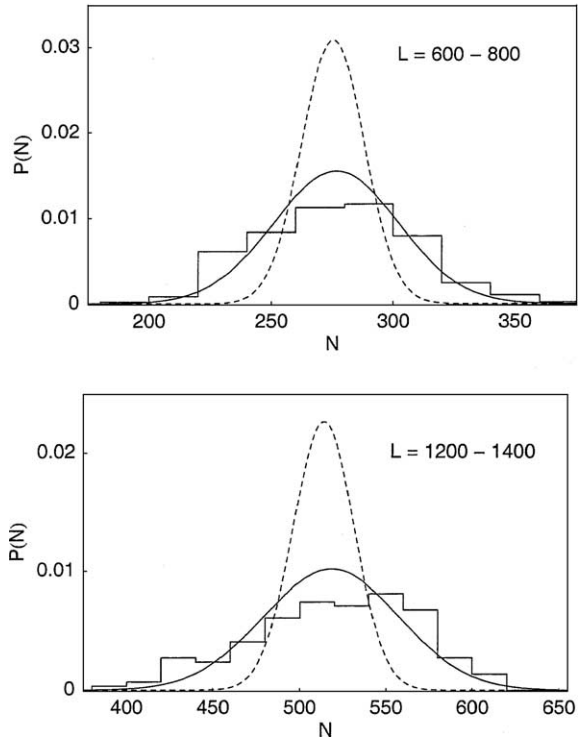

Fig. 10. Histograms giving the distribution of N values for the bands of L-values indicated in Figs. 8 and 9 with $\langle L\rangle$=700 and 1300 respectively. The windows for the N variable are 20 bases. The solid curves are plots of the distribution function of Eq. (26) using the power law σ_L* of Eq. (28). The dashed curves are plots of Eq. (26) using the random distribution σ'_L of Eq. (27).

L=1200–1400 are indicated in Figs. 8 and 9. We then simply count how many of the genes having L values in these ranges have N values that fall within given ranges. We take the size of the windows of N values as ΔN=20 bases. We then construct a histogram based on the fraction of N values falling in a given window. The resulting histrogram approximations to the distributions so constructed for the L bands illustrated in Figs. 8 and 9 are shown in Fig. 10. The solid curves in Fig. 10 give plots of Eq. (26) using σ of Eq.

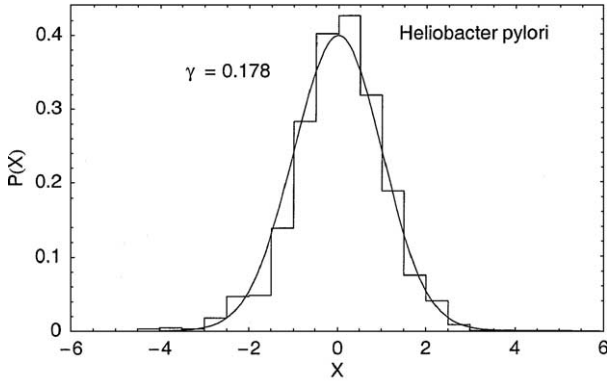

Fig. 11. Histogram of the $P(X)$ distribution for *Heliobacter pylori* based on the data of Fig. 9. The windows for the X variable are ΔX=0.5. The solid curve is a plot of the simple Gaussian distribution of Eq. (31).

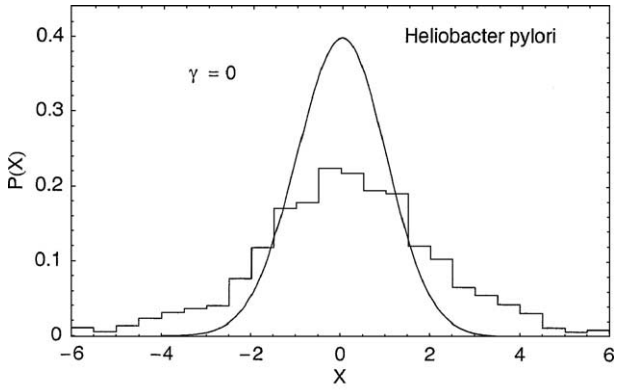

Fig. 12. Histogram of the $P(X)$ distribution for *Heliobacter pylori* based on the data of Fig. 7 converted to the X_i variables of Eq. (32) using the value of γ=0 for the random distribution. The solid curve is a plot of the simple Gaussian distribution of Eq. (31).

(28) for the power law case with $\langle L\rangle$=700 and 1300 respectively for the two cases. The dashed curves in Fig. 10 give plots of Eq. (26) using σ of Eq. (27) for the random case for the same two values of $\langle L\rangle$. There are only a limited number of points in the L bands shown in Fig. 8 and so the histograms are a bit ragged, but clearly they are much closer to the power law distribution than the random distribution.

We return to the scaled data given in Fig. 9. To test further if the power law scaling given in Eq. (32) indeed reduces the gene distribution to the simple Gaussian of Eq. (31), we construct a histogram for all of these data using box sizes of ΔX=0.5 using the value of gamma given in Eq. (12), namely γ=0.178. The resulting histogram is shown in Fig. 11 where the solid curve is a plot of the distribution given in Eq. (31). The agreement between the histogram and the continuous Gaussian distribution is excellent indicating that power law scaling of Eq. (32) is valid for genes as well as m-boxes.

If one was to use the σ for a random distribution as given in Eq. (27), then one would obtain the histogram for the Hp gene data shown in Fig. 12 where again the solid curve is the simple Gaussian of Eq. (31). In this case, the agreement between the histogram and the Gaussian distribution is very poor indicating that the random model is not a good for the C–G distribution of the Hp genes. In Fig. 10, the random distribution is seen to be much sharper than the power law distribution while in Fig. 12 random scaling gives a curve that is much broader than the power law distribution. This effect arises since in the scaling of Eq. (29) to give X one divides by sigma. Because the value of sigma for the random distribution is much smaller than that for the power law distribution, on dividing by this quantity, one produces an incorrectly scaled distribution that is too broad.

4. Distributions from moments

In addition to constructing histograms as approximate distributions for the C–G content of genes, one can

construct a continuous approximate distribution based on moments. Each point in the scatterplot of Fig. 9 is given by

$$X_i = \frac{\sqrt{L_i}(f_i - \bar{f})}{AL_i^{\gamma}} \tag{33}$$

We weigh each point given in Fig. 9 the same which is equivalent to giving each a probability of

$$P_i = 1/\mathrm{NG} \tag{34}$$

where we recall that NG, as given by Eq. (16), is the number of genes in the genome (1489 in Hp). The first two moments of the X distribution are then given by

$$\mu_1 = \frac{1}{\mathrm{NG}}\sum_{i=1}^{NG} X_i \quad \text{and} \quad \mu_2 = \frac{1}{\mathrm{NG}}\sum_{i=1}^{\mathrm{NG}} X_i^2 \tag{35}$$

We then use these quantities to construct the distribution

$$P(X)\mathrm{d}X = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\left(X - \bar{X}2\right)/2\sigma_X^2\right]\mathrm{d}X \tag{36}$$

where σ_X is given in terms of μ_1 and μ_2 by Eq. (3) and $\bar{X}=\mu_1$. The results of this construction are shown in the upper graph in Fig. 13. The solid dots represent the values of the standard Gaussian of Eq. (31) while the solid curve is the distribution given in Eq. (36) constructed from the moments given in Eq. (35) obtained by summing over all of the points given in Fig. 9. One sees that the two distributions are virtually identical thus indicating that the power law scaling of Eq. (28) is extraordinarily accurate for the C–G distribution for the genes in Hp.

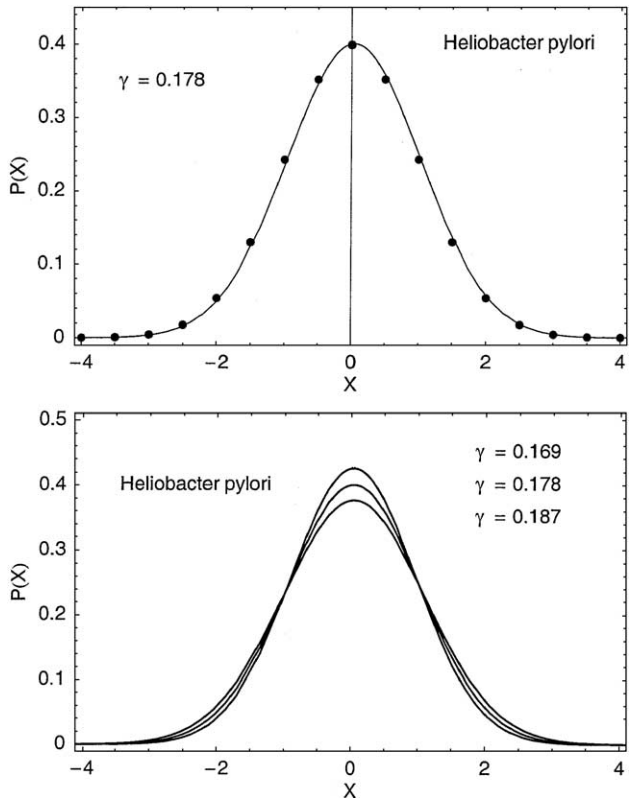

Fig. 13. Upper graph—the solid curve is the distribution $P(X)$ of Eq. (36) constructed from the moments of the X_i distribution given in Fig. 9. The solid points represent the values of the simple Gaussian distribution of Eq. (31). Lower graph—the distribution $P(X)$ of Eq. (36) constructed from moments of the X_i data obtained from Eq. (33) for the three values of γ indicated.

To illustrate the sensitivity of the distribution function to the correct value of the persistence exponent γ, we have repeated the moment calculation outlined above using sets of the X_i variables given by Eq. (33) for values of γ that are 5% higher and lower, respectively, than the value of γ=0.178 for Hp given by Eq. (12). The results of these calculations are shown in the lower graph in Fig. 13 and one sees that even a small variation in the value of γ makes a big difference in the goodness of fit of the resulting distribution function.

5. Comparison of scaled distributions

The scaling relation given in Eq. (32) translates the set of length (L_i) and C–G content (N_i) variables for the genes of a given species into a set of variables X_i the distribution function for which is given by the simple Gaussian distribution of Eq. (31). The scaling relation of Eq. (32) requires the parameters γ, A, and $\bar{f}$ that in general will be different for each species. We have seen in Figs. 11 and 13 that this scaling function is extremely accurate for the genes of *Heliobacter pylori*. In this section, we want to apply the same procedure to four other species to test whether Eq. (32) indeed is a universal scaling relation.

We examine four species that we have used as examples previously [1,2]. The species chosen are *Mycoplasma pneumoniae* [6], *Treponema pallidum* [7], *Thermoplasma volcanium* [8], *and Rickettsia prowazekii* [9]. The genomes of all of these species can be obtained from institute for genomic research (Tigr) on the web [4]. All of the species treated are bacteria except for *Thermoplasma volcanium* which is an archaean. For all of the species treated, the complete genome is a single circular chromosome. The essential parameters for the genomes of these species are listed in Table 1.

The analysis of the C–G content distribution for these species follows that outlined previously for Hp. One first constructs the analogs of Fig. 9 using Eq. (32). One can then construct a histogram approximation for the distribution and the Gaussian distribution of Eq. (36) based on the moments of the X_i set as outlined in Eq. (35). One then compares these approximations of the distribution to the simple Gaussian distribution of Eq. (31). If the scaling of Eq. (32) is valid, the approximate distributions should coincide with the simple Gaussian distribution.

The results of this process are shown in Fig. 14 for *Treponema pallidum* and *Rickettsia prowazakii* and in Fig. 15 for *Thermoplasma volcanium* and *Mycoplasma pneumo-*

Table 1
Genome parameters

Species	N	NG	f_{cg}	$\bar{f}$	A	γ	D	GL
Hp	1,643,831	1489	0.390	0.396	0.303	0.178	1.322	1104
Mp	816,394	688	0.399	0.403	0.203	0.297	1.203	1187
Rp	1,111,523	842	0.289	0.300	0.693	0.075	1.425	1320
Tp	1,138,012	1039	0.527	0.529	0.149	0.318	1.182	1095
Tv	1,584,804	1494	0.399	0.409	0.161	0.290	1.210	1061

Species: Hp—*Heliobacter pylori*; Mp—*Mycoplasma pneumoniae*; Rp—*Rickettsia prowazekii*; Tp—*Treponema pallidum*; Tv—*Thermoplasma volcanium*.
Abbreviations: N—total number of bases in the complete genome; NG—number of genes; f_{cg}—fraction of C–G base pairs in complete genome; $\bar{f}$—average fraction of C–G bases in the genes; A—constant in scaling law; γ—power law exponent; D—$3/2-\gamma$=fractal dimension; GL—average gene length.

niae. In each case, the solid curve is the simple Gaussian distribution of Eq. (31) while the dashed curves are the Gaussian distribution of Eq. (36) based on the moments of the X_i set. The box curves are the histograms (with window width $\Delta X=0.5$) based on the analogs of Fig. 9. One sees that indeed the scaling of Eq. (32) does reduce all of the gene sets $\{L_i,\ N_i\}$ closely to the same simple Gaussian distribution of Eq. (31). From Table 1, one sees that the values of γ for the species treated vary over a wide range. We note that the variation we see in Figs. 14 and 15 is within the range of variation shown in the lower graph in Fig. 13 for 5% variation in γ for Hp.

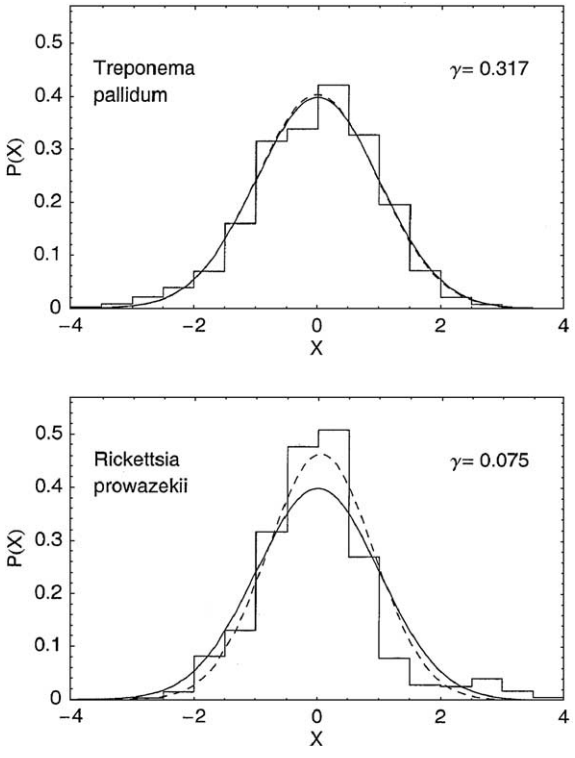

Fig. 14. Histogram of the $P(X)$ distribution for the species indicated based on data analogous to that shown in Fig. 9. The windows for the X variable are $\Delta X=0.5$. The solid curve is a plot of the simple Gaussian distribution of Eq. (31) while the dashed curves are the distribution $P(X)$ of Eq. (36) constructed from the moments of the appropriate X_i set.

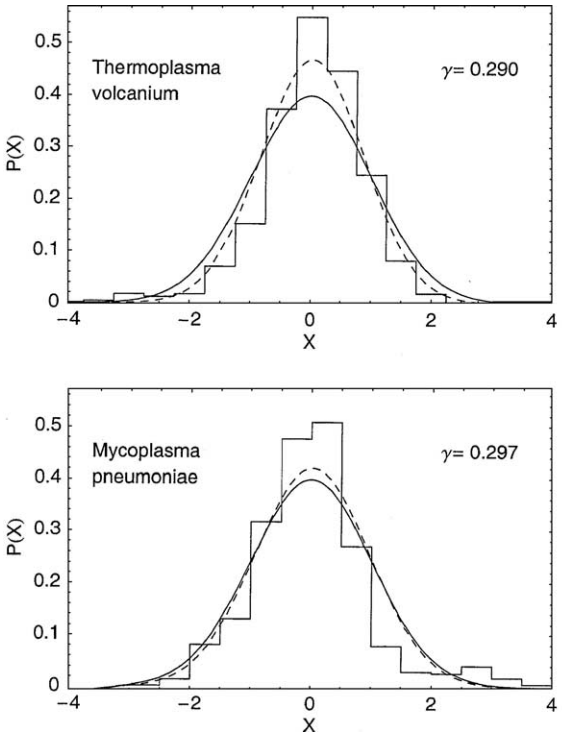

Fig. 15. Histogram of the $P(X)$ distribution for the species indicated based on data analogous to that shown in Fig. 9. The windows for the X variable are $\Delta X=0.5$. The solid curve is a plot of the simple Gaussian distribution of Eq. (31) while the dashed curves are the distribution $P(X)$ of Eq. (36) constructed from the moments of the appropriate X_i set.

As a final comparison, we take the distribution functions bases on moments for all five species listed in Table 1 and combine them in a single plot. This is shown in Fig. 16 where the solid dots give the values of the simple Gaussian of Eq. (31). The resulting tight package of curves indicates that the scaling of Eq. (31) indeed represents a universal scaling of the C–G distribution in genes.

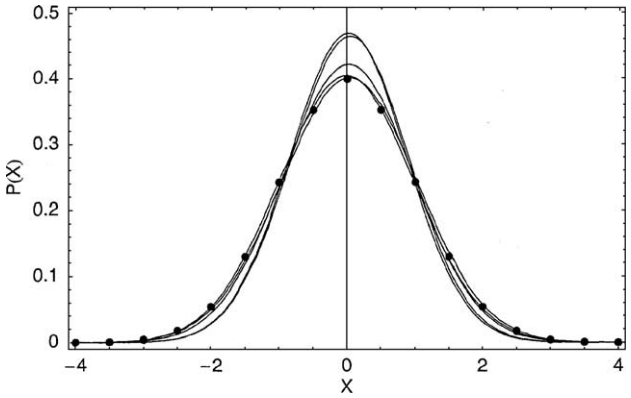

Fig. 16. A compilation of the distributions $P(X)$ of Eq. (36) constructed from the moments of the appropriate X_i set for the five species listed in Table 1. The solid points represent the values of the simple Gaussian distribution of Eq. (31).

References

[1] D. Poland, The persistence exponent of DNA, Biophys. Chemist. 110 (2004) 59–72.
[2] D. Poland, The phylogeny of the persistence exponent of DNA, Biophys. Chemist. 112 (2004) 233–244.
[3] R.A. Alm, et al., Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Heliobacter pylori*, Nature 397 (6715) (1999 (Jan. 14)) 176–180.
[4] The worldwide web address of The Institute for Genomic Research is: http://www.tigr.org.
[5] B.B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman and Company, New York, 1982.
[6] R. Himmerlreich, et al., Complete sequence analysis of the genome of the bacterium *Mycooplasma pneumoniae*, Nucleic Acids Res. 24 (22) (1996 (Nov. 15)) 4420–4449.
[7] C.M. Fraser, et al., Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, Science 281 (5375) (1998 (Jul. 17)) 375–388.
[8] T. Kawashima, et al., Archael adaption to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*, Proc. Natl. Acad. Sci. U. S. A. 97 (26) (2000 (Dec. 19)) 14257–14262.
[9] S.G. Anderson, et al., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, Nature 396 (6707) (1998 (Nov. 12)) 133–140.