A Universal Signature in Whole Genomes

Ta-Yuan Chen*, Li-Ching Hsieh*‖, Chang-Heng Chang*, and H.C. Lee*†‖

*Department of Physics, †Department of Life Sciences and ‖Center for Complex Systems, National Central University, Chungli, Taiwan 320.

*To whom correspondence should be addressed E-mail: hclee@phy.ncu.edu.tw

Supporting Online Material

Materials and Methods

Complete genome sequences

Complete sequences of the 155 prokaryotes are taken from GenBank: http://www.ncbi.nlm.nih.gov/genomes/Complete.html (2003 December 11) and the 127 complete chromosomes of 10 eukaryotes are taken from http://www.ncbi.nlm.nih.gov/genomes/static/euk_g.html (2003 May 31). The ten eukaryotes (common name and number of chromosomes in brackets) are *A. thaliana* (mustard, 5), *C. elegans* (worm, 6), *D. melanogaster* (fruit fly, 6), *E. cuniculi* (11), *H. sapiens* (human, 24), *M. musculus* (mouse, 21), *P. falciparum* (malaria parasite, 14), *R. norvegicus* (rat, 21; Chromosome Y missing), *S. cerevisiae* (yeast, 16) and *S. pombe* (fission yeast, 3). Fig. S1 gives an indication of the diversity of the complete sequences in terms of sequence length L and fractional A+T content p among all sequences.

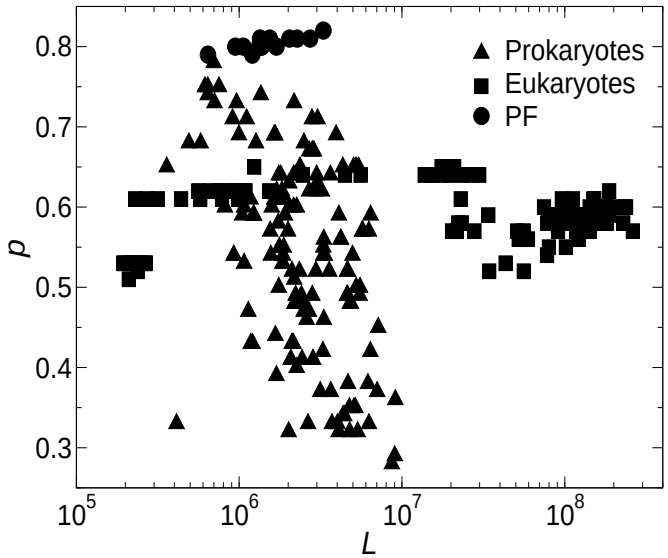

Figure S1. Diversity of complete sequences in sequence length L and fractional A+T content p. Each symbol in the plot represents a complete sequence, triangles for the 155 prokaryotes and squares for the 113 eukaryotes. The 14 *P. falciparum* (PF) chromosomes (bullets) have the largest p.

Generation of k-spectrum

Occurrence frequencies of the 4^k k-mers are determined by sliding a k-nucleotide-wide window one nucleotide at a time across the genome and registering the number of times each k-mer appears in the window. This generates a complete set of occurrence frequencies, which is then converted to the set $\{n_f | f = 1, 2, \ldots\}$, called a k-spectrum in the main text, where n_f is the number of k-mers occurring with frequency f in the genome. Passing the sliding window once over the genome suffices to generate the sets for all k's (*1*).

Computation of reduced spectral width

The reduced spectral (RdSW) width is given by

$$\mathcal{M}_\sigma \equiv \sum_{m=0}^{k} w_{k,m}(\sigma_m/\sigma'_m)^2; \qquad w_{k,m} = L^{-1}2^k(k,m)\bar{f}_m(p), \tag{1}$$

where σ_m is the relative spectral width (RlSW; standard deviation of mean frequnecy) computed from the m-set subspectrum of the genome, $\sigma'_m{=}(1-2^{1-k})/\bar{f}_m(p)$ is the RlSW expected of the subspectrum of the

random copy, $\bar{f}_m(p)=\bar{f}2^k p^m(1-p)^{k-m}$ is the mean frequency of the m-set, (k,m) is the binomial satisfying $\sum_m(k,m)p^m(1-p)^{k-m}=2^k$ and $\bar{f}=L/4^k$ is the overall mean. Note that the weights $w_{k,m}$ satisfy the relation $\sum_m w_{k,m}=1$. In the case $p=0.5$, the m-set subspectra collapse to a single k-spectrum of half-width σ, $\bar{f}_m(p)\approx\bar{f}$ and we have

$$\mathcal{M}_\sigma=(\sigma/\sigma')^2, \qquad (p=0.5) \tag{2}$$

where $\sigma'=(1-2^{1-k})/\bar{f}$ is the expected RlSW of the k-spectrum of the random copy.

Data sets on effective root-sequence lengths

For a given k the effective root-sequence length $L_r(k)$ for each genome sequence is defined as $L/\mathcal{M}_\sigma$, where L is the sequence length and $\mathcal{M}_\sigma$ is the RdSW of the k-spectrum of the sequence. The $L_r(k)$'s were observed to be k-dependent but substantially genome independent. Tables S1-S3 give the $L_r(k)$'s averaged over sets of sequences. Table S1 gives the $L_r(k)$'s for four sets of genome sequences: the 155 prokaryotes (PK); the 113 eukaryotes (EK) not including *P. falciparum*; the noncoding regions of PK; the PK and EK combined (CB). Each sequence in the "noncoding PK set" is obtained by concatenating all positively oriented genes in both strands of a genome. The PK, EK and noncoding PK sets of data are shown as black symbols in Fig. 3 (A) in the main text. The CB set is shown as black triangles in Fig. 3 (B) in the main text.

Table S1. $L_r(k)$'s for four sets of genome sequences: the prokaryotes (PK); the eukaryotes (EK); the noncoding sequences in the prokaryotes; the prokaryotes and eukaryotes combined (CB). Entries after the "±" gives the standard deviations on $L_r(k)$.

k	Prokaryotes (PK)	Eukaryotes (EK)	noncoding PK	PK + EK (CB)
2	3.627 E2 ± 2.190 E2	2.495 E2 ± 1.559 E2	4.543 E2 ± 5.372 E2	3.149 E2 ± 2.028 E2
3	7.433 E2 ± 3.902 E2	6.051 E2 ± 2.607 E2	9.390 E2 ± 6.473 E2	6.851 E2 ± 3.484 E2
4	1.749 E3 ± 8.406 E2	1.602 E3 ± 6.334 E2	2.139 E3 ± 1.294 E3	1.687 E3 ± 7.638 E2
5	4.509 E3 ± 2.075 E3	4.381 E3 ± 1.682 E3	5.260 E3 ± 2.924 E3	4.455 E3 ± 1.920 E3
6	1.231 E4 ± 5.608 E3	1.221 E4 ± 4.657 E3	1.324 E4 ± 7.079 E3	1.226 E4 ± 5.229 E3
7	3.440 E4 ± 1.600 E4	3.240 E4 ± 1.301 E4	3.209 E4 ± 1.727 E4	3.356 E4 ± 1.484 E4
8	9.678 E4 ± 4.587 E4	7.942 E4 ± 3.593 E4	6.917 E4 ± 3.875 E4	8.946 E4 ± 4.283 E4
9	2.589 E5 ± 1.232 E5	1.682 E5 ± 9.807 E4	1.227 E5 ± 7.302 E4	2.206 E5 ± 1.218 E5
10	5.995 E5 ± 2.849 E5	3.126 E5 ± 2.517 E5	1.767 E5 ± 1.144 E5	4.785 E5 ± 3.062 E5

Table S2 gives the $L_r(k)$'s for three sets of test sequences: set of 115 random sequences (RN) matching the profiles of genomes in the PK set; set of 115 PK-matching replicas (RP) of random root-sequences of length 300 b; set of 268 CB-matching sequences generated by the model with the parameters L_0=1000, l_x=500 and r=0.33 (Model A). Note that the RN set does not form a universality class because for each random sequence $L_r(k)$ is approximately the sequence length. This is reflected in two aspects of the data: (a) the average is about 3 Mb, independent of k (except for k=2); (b) the large, k-independent standard deviations reflect the span in the lengths of genomes in PK, which range from 0.2 to 7 Mb. The RP is a trivial universality class with $L_r(k)$ being a constant approximately equal to the root-sequence length of 300 b. The Model A set is a nontrivial universality class that is not genome-like. The three sets of data are shown as green symbols in Fig. 3 (A) in the main text.

Table S2. $L_r(k)$'s for three sets of test sequences. See text for the a description of the sets.

k	Random sequences (RN)	Replicas (RP)	Model A
2	3.956 E6 ± 4.862 E6	3.875 E2 ± 2.441 E2	1.832 E3 ± 3.333 E2
3	2.787 E6 ± 1.783 E6	2.887 E2 ± 7.117 E1	2.410 E3 ± 6.453 E2
4	2.654 E6 ± 1.585 E6	2.787 E2 ± 3.591 E1	3.774 E3 ± 1.152 E3
5	2.636 E6 ± 1.549 E6	2.818 E2 ± 2.174 E1	6.164 E3 ± 1.764 E3
6	2.672 E6 ± 1.576 E6	2.835 E2 ± 1.350 E1	1.045 E4 ± 2.810 E3
7	2.704 E6 ± 1.597 E6	2.856 E2 ± 1.254 E1	1.783 E4 ± 4.660 E3
8	2.717 E6 ± 1.600 E6	2.862 E2 ± 1.378 E1	3.038 E4 ± 7.872 E3
9	2.725 E6 ± 1.606 E6	2.859 E2 ± 1.531 E1	5.153 E4 ± 1.360 E4
10	2.731 E6 ± 1.610 E6	2.854 E2 ± 1.688 E1	8.694 E4 ± 2.363 E4

L_r of *P. falciparum* and some other organisms that are parasitic, symbiotic and/or have genomes with "extreme" base compositions

The L_r's of *P. falciparum* (PF), an eukaryotic parasite that has the most biased base composition (p=0.81±0.11), are very distinct from those of all the organisms in CB. However, neither being parasitic nor being highly compositionally biased, nor being both, can be the cause for this distinctness. In Table S3 we give the L_r's of PF, averaged over 14 chromosomes, together with those of: *E. cuniculi*, another eukaryotic parasite (*2*); *Wigglesworthia glossinidia brevipalpis* (WG), a symbiont and the most AT-rich prokaryote (in PK) (*3*); *Buchnera aphidicola*, another AT-rich prokaryote (*4*); *Rickettsia prowazekii* an AT-rich, parasitic, mitochondria-like prokaryote (*5*); *Streptomyces coelicolor*, the largest and most GC-rich member in PK (*6*). Like PF, the genome of each of these five organisms is atypical in its own way. The data in Table S3 show that PF alone is clearly a class by itself. The anomalously large $L_r(2)$ of PF is partly caused by the fact that the chromosomes of PF are exceptionally close to being exactly compositionally self-complementary.

Table S3. $L_r(k)$'s for three sets of test sequences. See text for the a description of the sets.

k	PF (p=0.81)	*E. cuniculi* (p=0.53)	WG (p=0.78)	*B. aphidicola* (p=0.75)	*R. prowazekii* (p=0.71)	*S. coelicolor* (p=0.28)
2	1.505 E3	2.591 E2	2.829 E2	3.919 E2	1.431 E3	4.583 E2
3	2.943 E2	5.827 E2	4.590 E2	6.061 E2	2.495 E3	7.620 E2
4	3.972 E2	1.430 E3	9.387 E2	1.122 E3	4.890 E3	1.281 E3
5	5.664 E2	3.916 E3	2.080 E3	2.426 E3	1.066 E4	2.895 E3
6	8.555 E2	1.114 E4	4.697 E3	5.575 E3	2.503 E4	7.128 E3
7	1.262 E3	3.042 E4	1.104 E4	1.356 E4	6.040 E4	1.633 E4
8	1.840 E3	7.303 E4	2.609 E4	3.341 E4	1.457 E5	4.138 E4
9	2.620 E3	1.344 E5	5.898 E4	7.924 E4	3.205 E5	1.096 E5
10	3.686 E3	1.834 E5	1.269 E5	1.720 E5	5.806 E5	2.697 E5

Generation of model genome sequences

The model has three explicit parameters: L_0, the initial sequence length; l_x, twice the average length of duplicated segments; r, point mutation (replacement only) rate per base. For simplicity the lengths l of the duplicated segments are selected using a square distribution with range $1 \le l \le l_x$. In our model p and 1-p sequences are mathematically equivalent. In order to generate a set of model sequences having profiles that approximate a target genome set, corresponding to a target genome sequence with profile L and p, the initial random sequence (of length L_0) is chosen from among possible values a base composition closest to $p' = \min(p, 1-p)$, grown by segmental duplication to a length just longer than L, then given rL single replacements with compositional bias p'. The profile, L_m and p_m, of the final model sequence will then have $L_m \gtrsim L$ and $p_m \approx p'$. The initial sequences are compositionally self-complementary but otherwise random. For the main universality class, the optimum value for L_0 is 8 and sequence of this length can only have p=0, 0.25, 0.5, 0.75 or 1.0, so the initial sequences were chosen to have p=0.25 or 0.5. Two measures were taken to shorten computation time, neither of which is expect to qualitatively affect the presented results. Firstly, because $L_0 << l_x$, an initial sequence was first replicated to a length just greater than l_x before it was subjected to growth by stochastic segmental duplication. Secondly, for model eukaryote sequences, l_x was changed to 10000 once the sequence grows beyond 2 Mb.

Chi-squared search for optimum parameters in growth model

Let $\mathcal{G}$ be a target set of N genome sequences whose averaged L_r's are given by the data set $\{L_r\}_\mathcal{G} \equiv \{L_r(k) \pm \delta(k) | k = 2, \cdots, 10\}_\mathcal{G}$. We want to find a set $\mathcal{M}$ of N model sequences with profiles matching the sequences in $\mathcal{G}$ whose set of individual L_r's, $\{L_r\}_\mathcal{M} \equiv \{L_{ri}(k) | i = 1, \cdots, N; k = 2, \cdots, 10\}_\mathcal{M}$, where i indexes the sequences, best agrees with $\{L_r\}_\mathcal{G}$. We define the chi-squared for $\mathcal{M}$ as

$$\chi^2_\mathcal{M} = \frac{1}{9N} \sum_{i=1}^{N} \sum_{k=2}^{10} \left(\frac{L_{ri}(k) - L_r(k)}{\delta(k)} \right)^2 . \tag{3}$$

For given $\mathcal{G}$, $\chi^2_\mathcal{M}$ is a function of the model parameters L_0, l_x and r. The optimum model parameters reported in the text were obtained by setting $\mathcal{G}$=PK and searching for the smallest $\chi^2_\mathcal{M}$ in the parameter space defined by L_0=8, 10 and $20 \le L_0 \le 200$ in steps of 20, l_x=20, 50, 100, 250, 500, 1000, 2000 and 4000 and $0 \le r \le 1.3$ in steps of 0.05. The results are shown in the eight panels in Fig. S2. Each panel is a color-coded χ^2 contour plot on an $r - L_0$ plane with fixed l_x. The color code for $\log \chi^2$ is given by the color bar shown to the right of the panels. We set $\mathcal{G}$ to PK rather than to the larger set CB because the difference between the $\{L_r\}_\mathcal{G}$'s for

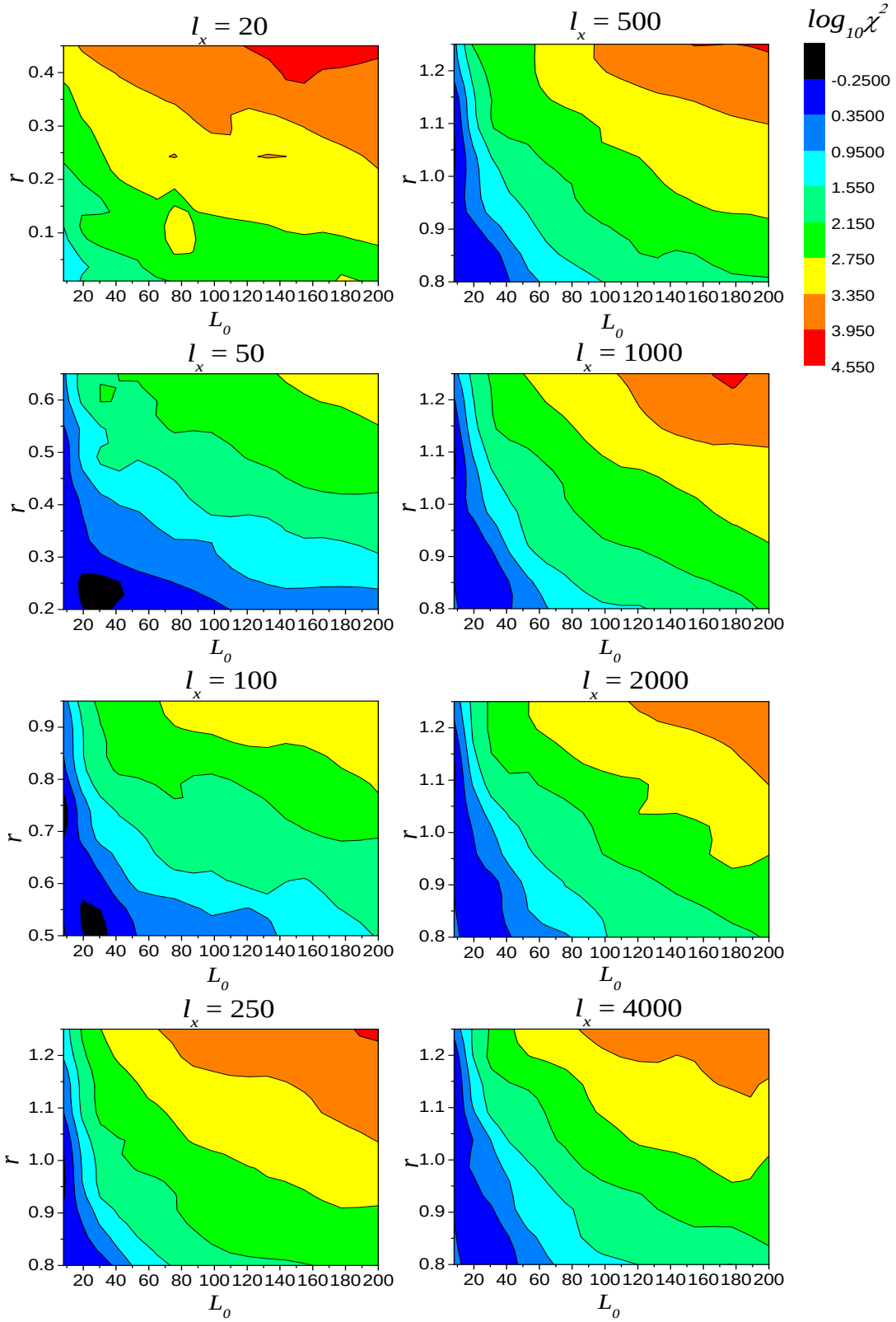

Figure S2. Color-coded contour plots of χ^2, defined in Eq. (3), which measures the closeness of L_r of sequences in a PK-matching set of model sequences to the L_r averaged over the PK set. Each panel gives the χ^2 as a function of r and L_0 at a fixed value of l_x. L_0, r and l_x are parameters of the model. Color bar to the right gives the color-code for $\log \chi^2$.

the two set is not statistically significant (see Table S1), and because, owing to the much greater lengths of the eukaryotic sequences (see Fig. S1), the computation time expended for the search for optimum parameters would have increased many fold had we set $\mathcal{G}$ to CB.

Generally, the optimum r and l_x are correlated: a larger l_x requires a larger r. Two regions of optimum parameters are seen, one characterized by $L_0 \approx 30$ and other by $L_0 \approx 8$. For this paper we reject the first region because the mutation rate appears to be too low, and also because the model sequences generated with this

set of parameters have too many repeats longer than 20 b (this aspect of the genome is not discussed in the main text). For the second optimum region (defined by $L_0 \approx 8$), it is seen that the χ^2 landscape is very steep in the L_0 direction but relatively gentle in the r and l_x directions. This implies that within the context of our model L_0 needs to be small, certainly far less than 200.

Results from model sequences

Table S4 gives the $L_r(k)$'s from CB, the 14 PF chromosomes and two sets of model sequences: a model CB set of 268 sequences matching the profiles of genomes in CB, generated with the parameters L_0=8, l_x=500, r=1.05; a model PF set of 14 sequences matching the profiles of the 14 PF chromosomes generated with the parameters L_0=80, l_x=500, r=0.24. The genome and model data are shown as black and green symbols, respectively, in Fig. 3 (B). Notice that the standard deviations on L_r of the model CB sequences are significantly smaller than their genomic counterparts. This implies that had we used a range of values for l_x and r centered around their respective optimum values of l_x=500 and r=1.05 (recall that their optimum values are correlated), we could have obtained a set of model sequences whose L_r have a spread more closely resemble the spread in the $L_r(k)$'s from CB.

Table S4. $L_r(k)$'s for the four sets: CB, model CB, PF; model PF. See text for the a description of the sets.

k	CB	Model CB	PF	Model PF
2	3.150 E2 ± 2.028 E2	5.966 E2 ± 5.010 E2	1.506 E3 ± 3.411 E2	6.178 E2 ± 6.654 E2
3	6.851 E2 ± 3.484 E2	9.975 E2 ± 1.883 E2	2.944 E2 ± 2.114 E1	3.596 E2 ± 1.442 E2
4	1.687 E3 ± 7.637 E2	2.079 E3 ± 3.268 E2	3.973 E2 ± 2.710 E1	4.176 E2 ± 1.055 E2
5	4.455 E3 ± 1.920 E3	4.795 E3 ± 7.632 E2	5.664 E2 ± 4.372 E1	5.368 E2 ± 9.020 E1
6	1.226 E4 ± 5.229 E3	1.206 E4 ± 2.115 E3	8.555 E2 ± 7.672 E1	7.431 E2 ± 8.064 E1
7	3.356 E4 ± 1.484 E4	3.195 E4 ± 6.357 E3	1.262 E3 ± 1.280 E2	1.062 E3 ± 8.446 E1
8	8.946 E4 ± 4.283 E4	8.699 E4 ± 2.005 E4	1.840 E3 ± 2.097 E2	1.551 E3 ± 9.357 E1
9	2.206 E5 ± 1.218 E5	2.304 E5 ± 6.515 E4	2.620 E3 ± 3.321 E2	2.283 E3 ± 1.267 E2
10	4.785 E5 ± 3.062 E5	5.915 E5 ± 2.166 E5	3.686 E3 ± 5.132 E2	3.389 E3 ± 1.976 E2

Time-dependence of accumulatine event rates

Suppose the instantaneous event rate per unit length per unit time, R_0, is a constant. Denote by $\langle R \rangle$ the accumulative rate averaged over the entire life span T of the genome and by R' the accumulative rate averaged over its late period from (t=) $T - \Delta T$ to T. We have T=4 (By), $\Delta T \approx 0.06$ and $\langle R \rangle / R' \approx 1/8$. Assume the genome length grew exponentially as

$$L(t) = L_0 \, e^{t/\tau} \tag{4}$$

This expression may be viewed as an approximation of

$$L(t) = aL_0 \, e^{t/\tau}/(a + e^{t/\tau})$$

for a range of t such that a is much greater than $e^{t/\tau}$. At some time t the number of events occurring in the interval dt is $dN(t) = R_0 L(t) dt$. The accumulative number of events from t=0 to t is

$$N(t) = R_0 L_0 \int_0^t e^{t/\tau} dt = \tau R_0 L(t)(1 - e^{-t/\tau})$$

and the accumulative rate per length is

$$R(t) = N(t)/(tL(t)) = (\tau R_0/t)(1 - e^{-t/\tau}) \tag{5}$$

Then at t=T

$$\langle R \rangle = N(T)/(L(T)T) = (\tau R_0/T)(1 - e^{-T/\tau}) \approx \tau R_0/T \qquad (\tau << T) \tag{6}$$

The accumulative number of events over the late period ΔT is

$$\Delta N = \tau R_0 L(T)(1 - e^{-\Delta T/\tau})$$

so that

$$R' = \Delta N/(L(T)\Delta T) \approx R_0 \qquad (\Delta T << \tau) \tag{7}$$

That is, R' is just the terminal rate which is itself the constant rate R_0. Hence

$$\langle R\rangle/R' \approx \tau/T \approx 1/8 \tag{8}$$

With T=4 By and $L(T)$=3 Bb, we have $\tau\approx$0.50 By and $L_0\approx$1.0 Mb. It can be verified that given the value of the ratio in Eq. (8), the approximations taken in Eqs. (6) and (7) are in fact highly accurate.

References

1. B.L. Hao, H.C. Lee and S.Y. Zhang, Fractal related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*, **11**, 825-836 (2000).
2. See http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=6035
3. Akman, L. *et al.* Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat. Genet.* **32**, 402-407 (2002).
4. Tamas, I. *et al.* 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376-2379 (2002).
5. Andersson, S.G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondri. *Nature* **396**, 133-140 (1998).
6. Bentley, S.D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141-147 (2002).