Communication Model of DNA Replication

Bo Deng[1]

Abstract: The all-purpose communication model proposed by the author previously for DNA replication suggests that if the base pairing time of GC bases is between 1.65 and 3 times the base pairing time of AT bases (with the proportional ratio denoted by α), then the quaternary DNA system has the optimal information rate for all species on average. In the absence of a definitive experimental determination of the base pairing time ratio α, previous theoretical estimates based on hydrogen bonding energy put its range near $1.6667,\ 1.8268,\ 2$, all inside the optimal α interval $(1.65, 3)$ of the quaternary system. In this paper, we attempt to gauge the theoretical range alternatively. The basic assumption is that organisms which have an intrinsic capability to replicate the most information at base-4's optimal α ratio should be the most successful by evolutionary considerations. *Pelagibacter ubique* and *Prochlorococcus* are believed to be such organisms. Here we show that if these marine bacteria are capable of replicating the most information, then their α ratios indeed lie inside base-4's optimal interval.

Introduction. An organism has the same amount of genomic information when it is alive and dead. The difference is replication. The DNA code and its replication are two inseparable facets of cellular life. Each must have left telltale marks on the other through evolution. In this paper we take the view that replication exerts the foremost evolutionary pressure on species genomes because it operates at a time scale immeasurably faster than that of natural selection. We aim to find such evolutionary imprints of replication and to formulate an optimization theory by which its impact on the DNA code can be understood.

A communication model of DNA replication was proposed by the author in [4]. It treats species genomes as individual information sources but the DNA replication as an all-purpose communication channel when the DNA bases, $A,\ T,\ G,\ C$, are paired one at a time with their complementary bases along the single strands of the double helix. By this conceptual model, a cell can be thought as a receiver when it is newly formed and a transmitter when it is to duplicate. That however good a transmitter or receiver is for a communication system is not as critical as the system's channel which defines a definitive time bottleneck for information transmission.

Any Internet connection, such as dial-up, DSL, cable, optic fiber, etc, is an all-purpose channel through which all types of information travel. An all-purpose channel is characterized by its mean information rate (in bits per second), which measures the best average the channel does for all information types. However, each signal type, such as video, audio, spam, computer virus, etc, has its own information rate which may be more or less than the mean rate. However fast, there is an upper limit that no information rate can exceed, and the limit is called *the channel capacity*. A particular piece of information source may happen to go through the channel at the capacity

[1]Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588. E-mail: bdeng@math.unl.edu

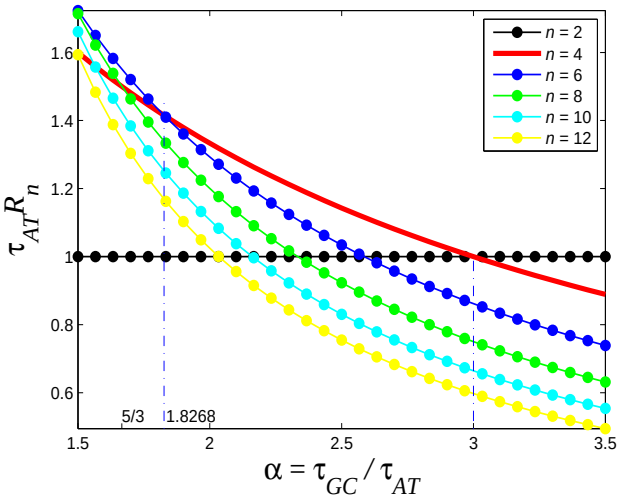


Figure 1: Discussed in detail in [4] are two critical values $\alpha = 5/3$ and $\alpha = 2$.

rate. Surprisingly, *any* information source can be *made* to go through the channel at a rate as close to the capacity rate as possible by properly encoding the source, one of the greatest discoveries by Claude E. Shannon's ([9]). However only the existence of such codes is guaranteed not the construction in general. Even if such a code is available, additional time is needed both before and after transmission to encode the source and to decode the signal respectively. Therefore, an information source that can naturally go through the channel at the capacity rate takes advantage of the channel the fullest.

The main finding of [4] is that if the pairing time of the hydrogen bonds of the GC pair is between 1.65 and 3 times that of the AT pair, then the information mean rate of the quaternary DNA system is greater than that of the binary model in either AT bases or GC bases alone, and greater than that of the model of any even bases. Thus, the result suggests that Nature may have favored the quaternary system because it can produce the best information rate for all species and on average, allowing the most species diversity to pass through the time bottleneck constrained by replication.

Of particular interest to this paper is the following simple mean rate model considered in [4]. Let n be the (even) number of bases of our communication model for DNA replication. Let $\tau_{_{AT}}$ be the base pairing time between the AT bases, $\tau_{_{GC}}$ be the base pairing time between the GC bases, and let $\alpha = \tau_{_{GC}}/\tau_{_{AT}}$ be the base pairing time ratio. Then the maximum information is $\log_2 n$ bits per base, and the mean replication rate is

$$R_n = \frac{\log_2 n}{\tau_{_{AT}}[1 + (\alpha - 1)(n - 2)/4]}$$

in bits per time. (See Appendix for a derivation of the formula.) Fig.1 shows the normalized rates $\tau_{_{AT}} R_n$ as functions of the base pairing time ratio α for a few even integers of base number n. It says the following. If $\alpha > 3$, then having the AT bases alone gives the best mean rate. On the other hand, if $\alpha \in (1.8268, 3)$, then having the quaternary system in $ATGC$ bases gives the

optimal mean rate. However, if $\alpha \in (1.5, 1.8268)$, then the mean rate would reach optimal with $ATGC$ bases plus a third hypothetical base pair. In other words, for each even base number n, the mean rate R_n is optimal only in an interval of the parameter α. Notice also that not all α values in base-4's optimal interval give the same mean rate R_4. It reaches the absolute maximum at the left end point $\alpha = 1.8268$ of the base-4 optimal interval.

The technical focus of this paper is instead on individual species which can naturally replicate at their channel capacity, also referred to as *the replication capacity* below. Such species make out the most that the quaternary replication machinery can offer, rushing through time the most genomic information ahead of other species.

The Result. Typically, an Internet channel has its own particular signal makeup in electrical waves or optical pulses, and its own number of signaling states referred to as symbols, say n of them. When an individual information source is encoded by the signal symbols for transmission, it results in a symbol frequency distribution, denoted by $p = \{p_1, p_2, \ldots, p_n\}$ with entries for symbol $1, 2, \ldots, n$ respectively. Let $\tau = \{\tau_1, \tau_2, \ldots, \tau_n\}$ denote the transmission times respectively for the 1st, 2nd, $\ldots$, nth signal symbols. Then for this source only, the kth symbol contains $H(p_k) = \log_2 1/p_k$ bits of information, and on average, each symbol contains $H(p) = p_1 \log_2 1/p_1 + p_2 \log_2 1/p_2 + \cdots + p_n \log_2 1/p_n$ bits for the source. Also on average and for this source only, each symbol takes $T(p, \tau) = p_1\tau_1 + p_2\tau_2 + \cdots p_n\tau_n$ units of time to transmit. Therefore, the particular information transmission rate for the source is $R(p, \tau) = H(p)/T(p, \tau)$, measured now in bits per unit time. $H(p)$ is called the *entropy* of the source, approximately measuring how diverse the source is per-symbol. It is a fact that $H(p) \leq H_n = \log_2 n$ for all p, and $H_n = H(p)$ if and only if the probability distribution p is the equidistribution: $p_k = 1/n$. That is, $H(p)$ reaches the maximum entropy, or per-base diversity, when each symbol is equally probable at every position of the transmitted signal. It is a trivial fact that the average or mean value of any probability distribution is the equidistribution: $(p_1 + p_2 + \cdots + p_n)/n = 1/n$. Thus, the channel's *mean information rate* is by definition $R_n(\tau) = H_n/T_n(\tau)$ in bits per unit time, where $T_n(\tau) = (\tau_1 + \tau_2 + \cdots + \tau_n)/n$ is the mean transmission time per symbol with the equiprobability assumption for the symbol frequencies. At the mean rate, the channel is maximal in transmitting the information content or per-base diversity, embracing all possible sources.

Although an information source's entropy after encoded in the signal symbols cannot exceed the channel entropy H_n, its source information rate $R(p, \tau)$ may be greater or smaller than channel's mean rate $R_n(\tau)$ depending on the symbol frequency p of the individual source because the source may take up very few or too many time-consuming signal symbols. By definition, the channel capacity is the maximum of $R(p, \tau)$ over all possible choices of the frequency distribution p, denoted by $K(\tau) = \max_p R(p, \tau)$. By a theorem from the Appendix, $K(\tau)$ is finite and the capacity-generating frequency satisfies

$$p_i = p_1^{\tau_i/\tau_1},\ \sum_{i=1}^{n} p_1^{\tau_i/\tau_1} = 1,\ K(\tau) = \frac{\log_2 1/p_1}{\tau_1} \tag{1}$$

Table 1: Base frequency distributions of various organisms

	Frequency							
Genomes	A	T	G	C	d	Δ_{AT}	$H(p)$	$\tau_{AT}R(p)^{**}$
S. coelicolor	13.9	14.0	36.1	36.0	0.2%	-44.2%	1.85	1.16
E. coli K-12	24.6	24.6	25.4	25.4	0.0%	-1.6%	1.99	1.41
E. coli O15:H7	24.8	24.7	25.2	25.2	0.1%	-1.0%	1.99	1.41
Human*	29.4	29.7	20.5	20.4	0.4%	18.2%	1.98	1.40
P. ubique	35.3	35.0	14.9	14.8	0.4%	40.6%	1.87	1.51
W. glossinidia	38.8	38.7	11.2	11.3	0.2%	55.0%	1.76	1.49

* The A, T, G, C contents shown are those of the chromosome #14 which has the greatest deviation from the generalized Chargaff law with $d = 0.4\%$ of all 23 chromosomes.

** $\alpha = 1.8268$ is used for the rate R of all genomes.

or equivalently the same equations replacing p_1, τ_1 by any fixed pair p_j, τ_j for any j.

For our communication model of DNA replication, $\tau_{_A} = \tau_1,\ \tau_{_T} = \tau_2,\ \tau_{_C} = \tau_3,\ \tau_{_G} = \tau_4$; $\tau_1 = \tau_2,\ \tau_3 = \tau_4$; and $\tau_{_{GC}} = \alpha\tau_{_{AT}}$ with $\tau_{_{GC}} = \tau_{_G} = \tau_{_C}, \tau_{_{AT}} = \tau_{_A} = \tau_{_T}$. Using the general result above, the capacity-generating base frequency satisfies:

$$\begin{aligned} &p_{_T} = p_{_A}^{\tau_{_T}/\tau_{_A}} = p_{_A},\ p_{_G} = p_{_C} = p_{_A}^{\tau_{_{GC}}/\tau_{_{AT}}} = p_{_A}^{\alpha} \\ &p_{_A} + p_{_T} + p_{_G} + p_{_C} = 2(p_{_A} + p_{_A}^{\alpha}) = 1 \\ &K(\tau) = \frac{1}{\tau_{_{AT}}}\log_2\frac{1}{p_{_A}} \end{aligned} \tag{2}$$

Table 1 shows the base frequencies of some selected bacteria as well as the base frequencies of Human chromosome #14. Let $p = \{p_{_A}, p_{_T}, p_{_G}, p_{_C}\}$ denote the base frequencies of a single strand DNA of the double helix of a chromosome, and $\bar{p} = \{\bar{p}_{_A}, \bar{p}_{_T}, \bar{p}_{_G}, \bar{p}_{_C}\}$ denote the base frequencies of its complementary strand. Then by Watson-Crick's base pairing principle, $\bar{p}_{_A} = p_{_T}, \bar{p}_{_T} = p_{_A}$, and similarly for the GC pair. Because of the complementarity, we see immediately that the base entropy H and the replication rate R are all invariant with respect to the choice of single strands. That is, $H(p) = H(\bar{p})$ and $R(p) = R(\bar{p})$. The table introduces two more strand invariant measures about which detailed discussions follow later. They are: the *intrapair frequency distance*

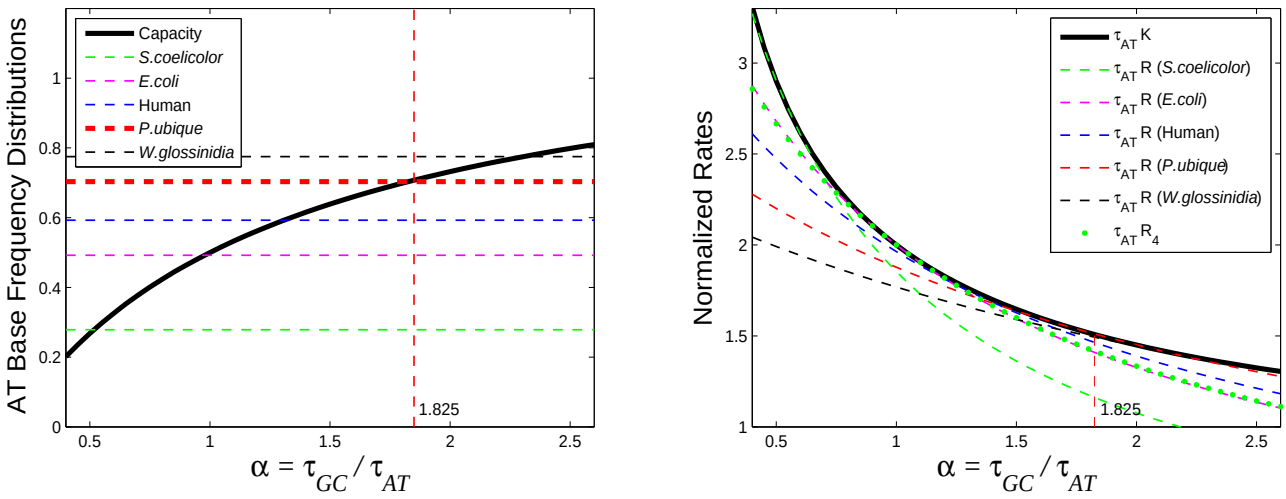
$$d = |p_{_A} - p_{_T}| + |p_{_G} - p_{_C}|,$$

and the *interpair frequency AT displacement*

$$\Delta_{AT} = (p_{_G} + p_{_C}) - (p_{_A} + p_{_T}).$$

Alternatively, with GC being the reference pair we have $\Delta_{GC} = -\Delta_{AT}$. The data are sorted in the increasing order of the AT content or Δ_{AT}.

Of 303 sequenced bacterial genomes ([11]), it shows the lest AT content for *Streptomyces coelicolor* ([11]); the most AT content for *Wigglesworthia glossinidia* ([1]); the first sequenced



(a) (b)

Figure 2: (a) For each pairing time ratio α, the capacity base distribution is solved from equations (2), and the joint $A+T$ frequency $p_{A+T} = 2p_A$ is plotted. (b) Comparison of the normalized replication capacity $\tau_{AT}K$, the individual replication rates $\tau_{AT}R$, and the mean diversity rate $\tau_{AT}R_4$.

organism *E. coli K-12* ([2]) and a relative strain *E. coli O15:H7* ([6]), both of which happen to have a near equidistribution of the bases. It also shows the base frequencies for *P. ubique* ([5, 10]) and the Human chromosome #14 ([12, 8]).

The single strand entropy column H of the table is self-explanatory. It is determined by organisms base frequencies. As for the replication information rate R, only the normalized rate $\tau_{AT}R$ is considered in this paper since individual species base pairing/replicating times may vary, (more discussions on the distinction follow below.) The last column of Table 1 shows the normalized rate for the α ratio at base-4's optimal value 1.8268. It shows that *P. ubique* has the best normalized replication rate.

Figure 2(a) shows the graphs of these organisms pairwise AT frequency: $p_{A+T} = p_A + p_T$, which is also strand invariant. It also shows the capacity-generating AT frequency curve as a function of the base pairing time ratio α, and how the curve crisscrosses organisms AT frequency lines. For example, the result implies that if $\alpha = 1.825$, then *P. ubique* replicates at the corresponding capacity rate. At one extreme, *C. michiganensis* is a capacity-replicating organism if $\alpha = 0.55$, and at the other extreme *W. glossinidia* is so if $\alpha = 2.35$. It depends on the ratio α. However, these two extreme α values lie outside the optimal mean rate range for the base-4 system.

In addition to mirroring the same information of Fig.2(a), Fig.2(b) shows the following. First, regardless the α value, no organism can replicate more information than the replication capacity K. Second, it shows that the slower the C and G bases pair with each other (larger α value), the smaller the replication capacity becomes, and the less frequent the GC pair should be in order to achieve the replication capacity. Third, it shows that the relationship between the replication rate

and the base distribution is not linear. For example, the human genome replicates more information per base than *W. glossinidia* does at $\alpha = 1.5$ but less information at $\alpha = 2$.

Discussion. *P. ubique* is considered to be one of the most successful organisms on Earth ([5]). It has the shortest genome that has the complete biosynthesis pathways for all 20 amino acids. It has no junk DNA. Its clad (SAR11) accounts for 25% of all microbial cells throughout the oceans. Our result above suggests that if *P. ubique* replicates at the information capacity, then its low GC to AT base content corresponds to a pairing time ratio α equal to 1.825, which is less than 0.1% difference from the optimal theoretical ratio 1.8268 predicted from the simplest communication model of DNA replication discussed in the Introduction. In fact, Fig.2(b) shows that *P. ubique* consistently has a greater normalized replication rate than others do for the range of $\alpha \geq 1.4$ which contains base-4's optimal mean rate range $(1.65, 3)$. Although there are alternative theories proposed for the low GC to AT content problem for bacteria in the oceans where nitrogen and phosphorous are frequently limited, they do not explain the problem for *P. ubique* according to [5].

Prochlorococcus is the smallest-known photoautotroph in the ocean whose biomass is similar to that of *P. ubique*. It is responsible for a significant fraction of global photosynthesis and carbon cycling ([3, 7]). A high-light-adapted strain (MED4) has an AT content, 69.2%, similar to that of *P. ubique*. The corresponding α value is 1.75, pushing even closer to the absolute optimal of base-4's optimal mean rate range $(1.65, 3)$. Taken together, their summed biomass in the ocean would spike against all other marine bacteria and the spike would be near the 70% AT-content range, which in turn corresponds to an α range near 1.8268. Interestingly, a low-light-adapted strain, MIT9313 ([7]), in deep sea has a significantly lower AT content (49.3%), and about 40% of its genes is not shared with its high-light-adapted counterpart ([7]).

Our communication model applies to single strand replication as well as to double strand replication. With respect to the single strand replication, the model, together with the complementary base pairing time assumption that $\tau_{_A} = \tau_{_T}, \tau_{_G} = \tau_{_C}$, leads to the following *generalized Chargaff law* (GCL):

$$p_{_A} = p_{_T},\ p_{_G} = p_{_C}$$

as seen in (2). This result is completely counterintuitive because it is the purine pair (AG) and the pyrimidine pair (TC) respectively that are similar in structure and elemental composition. However, the genome samples from Table 1 indeed support this GCL conclusion. We see now that the intrapair frequency distance d introduced in Table 1 measures a genome's deviation from GCL ($d = 0$). Contrasting the large variation in the interpair frequency displacement Δ_{AT}, the uniformity in the intrapair frequency distance d is striking. For the Human genome, for example, the largest discrepancy occurs on chromosome #14 (and #21), and the difference is no greater than 0.4% from the law. Of the 23 chromosomes, 9 satisfy GCL under the thousandth percentage point, and the rest are between 0.1% and 0.4%, see [8].

Applying our model to the double strand replication, it automatically recovers the classical Chargaff law ($d \equiv 0$). In fact, the double helical complementarity gives a *perfect* solution to the capacity optimization problem with $p_{_A} \equiv p_{_T}, p_{_G} \equiv p_{_C}$. This single-to-double Chargaff law pro-

gression is well in line with a well-known hypothesis that there was an RNA world before DNA evolved. More specifically, applying our communication model to single strand RNA replication leads to the GCL prediction. Hence, an argument can be made that GCL predated its double strand version, which only evolved later to give the perfect solution ($d \equiv 0$) to the capacity optimization problem. It can also be argued that the double helix structure of DNA is an evolutionary consequence of GCL rather than the other way around.

The homogeneity in the intrapair frequency distance ($d \sim 0$) for the replicating samples of Table 1 and the predicted GCL by the capacity theory together imply that all replicating organisms distribute their bases to achieve their own replication capacity. However, the heterogeneity in the interpair frequency displacement ($-44.2\% < \Delta_{AT} < 55\%$) implies that the pairing/replicating times τ_{AT}, τ_{GC} are species specific, with individual α ratios determined by the capacity frequency relation $p_{G} = p_{C} = p_{A}^{\alpha} = p_{T}^{\alpha}$. The normalized replication rate comparison suggests that if the *absolute* pairing/replicating time τ_{AT} for the AT pair were the same for all species, then *P. ubique* would have an *inherent* advantage over all others of Table 1 and for all α values from base-4's optimal mean rate interval.

For the mean rate replication model and for most of the preceding discussion on replication capacity, the base *pairing* times are assumed to be determined by the bonding energy of the hydrogen bonds of the nucleotides ([4]). In contrast, we did not gave a definitive definition to *cell replication time*. With the dichotomy of cells being transmitter and receiver, such a definition may aggregate some or all of the encoding and decoding times. In other words, it can be context dependent. Likewise, we did not consider the *absolute* pairing/replicating times τ_{AT}, τ_{GC}. They too can be context dependent because they may change not only throughout their evolutionary histories but also change with organisms developmental stages, or nutrient compositions, or ambient temperatures, etc. Our model did not nor could have taken into account all these context dependent variances. However, if one assumes that the absolute pairing/replicating times change proportionally with the same proportionality for both pairs from one given replication condition to another, then the time ratio $\alpha = \tau_{GC}/\tau_{AT}$ will remain constant for the changing replication conditions. As a result, the normalized information rate $\tau_{AT}R$ remains as a dimensionless constant, which in turn can be used for an intrinsic comparison among different species as we did here with Table 1 and Fig.2. This approach is analogous to that of [4] as well as Fig.1 where the normalized mean rate instead is used as an intrinsic comparison over replication models of different number of bases.

However context specific may it be, a particular definition of the base or cell replication times may have to be subjected to some more fundamental context-invariant rules. For instance, assuming that each species replicates at its own capacity, then the homogeneity in the intrapair frequency distance for the replicating samples of Table 1 again implies that whatever the definition of replication time may be, the moment that replication is considered completed is suggested by our model to be the moment when the hydrogen bonds of the complementary bases are paired, given rise to the pairing/replicating time symmetry: $\tau_{A} = \tau_{T} = \tau_{AT}, \tau_{G} = \tau_{C} = \tau_{GC}$. In other words, it is the base pairing time symmetry rather than a particular definition of replication times that leads

to the generalized Chargaff law. Further, it also suggests that it is by the same base pairing time symmetry rather than by chance, or base structure, or elemental composition, or individual fitness of natural selection that the generalized Chargaff law is satisfied.

Our model together with the empirical data from Table 1 implies that organisms settle down to their individual base pairing/replicating time τ_{AT}, τ_{GC}, which in turns determine their own α ratio, which in turn determines their base frequency distribution p according to Eq.(2). Thus, an argument can be made that each species genome is the result of its own information rate optimization.

The prevalence of intrapair distance homogeneity ($d \sim 0$) also suggests a base selection dominance over natural selection with respect to genomic composition. Think $d = 0$ as an evolutionary equilibrium, referred to as the *replication capacity equilibrium* below, for which the corresponding base frequency distribution is selected through replication optimization. Then, the empirical data suggests that the genomic composition that is due to evolution by natural selection falls inside a small vicinity ($0 \leq d < 0.4\%$) of the capacity equilibrium. More specifically, for a given base capacity distribution p, there are different base permutations to realize the equilibrium distribution. Thus, our result suggests that a species genome take up at any given time in its evolutionary history one particular realization in a small neighborhood of its replication capacity equilibrium.

Our model also provides an implicit, mechanistic explanation to this replication dominance. The mechanistic principle holds for any dynamical process which has two or more competing subdynamics operating at diverse time scales. That is, between a slow subdynamics and a fast subdynamics, the fast process always dominates — with the combined dynamics closely tracking the fast constituent's equilibrium. With regard to genomic evolution, DNA replication operates at a much faster time scale than natural selection does, with fractions of a second v.s. hundreds of thousand years, a practical order of infinity. As a result, the replication capacity equilibrium dominates. Hence, it can be argued that whenever natural selection forces a particular base distribution away from its replication equilibrium, the information rate optimizing force always brings the disturbance quickly back to a new realization of the equilibrium. As an interpretation to this replication v.s. evolution tug, a conjecture can be made that the greater deviation the intrapair distance is from its equilibrium $d = 0$, the more recent evolutionary changes took place. For example, with regard to the microbes from Table 1, this conjecture would apply to *P. ubique*, the Human chromosome #14, as well as chromosome #21 with $d = 0.4\%$, chromosome #16, #20, and the Y chromosome, all having the second largest $d = 0.3\%$.

This does not mean the capacity theory can replace the theory of natural selection. Quite to the contrary, it leaves a huge opening for natural selection to operate. More specifically, the theory does not address the question of species-specific ratio α, which leads to the heterogenous distribution in the interpair frequency displacement Δ_{AT}. This heterogeneity appears to correlate with species differentiation. Thus, a conjecture can be made that natural selection primarily impacts on the interpair frequency displacement Δ_{AT} while DNA replication primarily impacts on the intrapair frequency distance d.

With few exceptions, viruses do not self-replicate. Without the primary replication pressure,

Table 2: Base frequency distributions of various viruses

	Frequency							
Genomes	A	T	G	C	d	Δ_{AT}	$H(p)$	$H(P)$
phage VT2-Sa	25.6	24.5	26.9	23.0	5.0%	0.2%	1.9976	2.0000
phage 933W	27.6	22.8	27.4	22.2	10.4%	0.8%	1.9927	1.9999
phage P1	26.1	26.6	23.5	23.8	0.8%	5.4%	1.9978	1.9979
phage phiX174	24.0	31.3	23.3	21.5	8.1%	10.6%	1.9846	1.9921
phage T4	31.8	32.9	16.5	18.8	3.4%	29.5%	1.9355	1.9367

there is no reason to expect their genomes to track closely the replication capacity equilibrium $d = 0$ for self-replicating organisms. Table 2 indeed supports this observation. Surprisingly though the displayed entropy patterns can also be explained by the optimization paradigm proposed here. Without replication, the static single- and double-strand per-base diversity entropies reach the maximum $H_4 = 2.0$ bits per base when all bases are equally probable $p_{_A} = p_{_T} = p_{_G} = p_{_C} = 1/4$. As shown in Table 2, the entropies are indeed uniformly near the maximum even though d varies considerably, suggesting that the equiprobability is not a too stringent condition for the maximization. When both strand are taken together, the base frequencies are further homogenized as $P_{_A} = P_{_T} = (p_{_A} + p_{_T})/2,\ P_{_G} = P_{_C} = (p_{_G} + p_{_C})/2$ so that $d \equiv 0$ for the double-strand Chargaff Law. As a result, the aggregated double-strand per-base entropy $H(P)$ gets even closer to the 2 bits per-base maximum. Hence, an argument can be made that the primary function of virus genomes is to maximize their stationary per-base information entropy. Furthermore, the heterogeneity in both d and Δ_{AT} can be viewed as indicators of their ongoing evolutionary differentiations.

Genomic diversity is at least two dimensional, one is obvious and the other is not. The obvious is of the genome length of an organism. The not-so-obvious is the information entropy $H(p)$ each base carries for a base distribution p, which is length independent. The total information content of a genome of length L and of base distribution p is $H(p) \times L$. The replication rate is not about the genomic length nor the base entropy per se rather than the information entropy that is replicated in a unit time. In particular, it is sampling time invariant — i.e., the rate calculation results in the same value whether a time interval of one second or one hour is used by an observer. Given the same length of genomes, the work of [4] shows that the quaternary replication system gives the best mean information rate if $\alpha \in (1.65, 3)$. In contrast, our result here strongly suggests that *P. ubique* and *Prochlorococcus* have the best intrinsic rate for the same α range.

DNA code is unique in a fundamental way that it has to be constantly maintained and updated by replication. This is reflected by the ways how information is measured in our model. The measurement by the information entropy $H(p)$ is static whereas the measurement by the information rate $R(p)$ is dynamic. As an illustration, compare a genome of 50% AT content to a genome of 70% AT content but both having the same α value 1.825 as *P. ubique*'s. Then the former has a static entropy of 2 bits per base and a replication rate of 1.4125 bits per unit of AT

pairing time. Respectively, the latter has a static entropy of 1.8813 bits per base but a replication rate of 1.5081 bits per unit of AT pairing time. Thus, the 70% AT-content genome has about $(2 - 1.8813)/2 = 0.0594 \sim 6\%$ less static information than its 50% counterpart. This lost is accumulative only in length not in time. That is, a 50% AT-content genome of twice the length of a 70% counterpart has about $2 \times 0.0594 \sim 12\%$ more total information. In contrast, the 70% AT-content genome replicates about $(1.5081 - 1.4159)/1.4159 = 0.0651 = 6.5\%$ more information in each unit of AT pairing time than its 50% counterpart does. This gain is accumulative in time: in two units of AT pairing time, the 70% genome replicates $2 \times 6.5 = 13\%$ more information than the 50% genome does, and $3 \times 6.5 = 19.5\%$ more information in 3 units of AT pairing time, etc. This is due to the time dynamical nature of the information rate measurement. Multiplying this small gain by a factor of millions or billions of AT pairing time unit throughout their common evolutionary history, the net information gain is astronomical. We simply suggest here that this extra amount of information must translate in part into some greater evolutionary successes for the replicator of higher $R(p)$ rate. In other words, as far as information is concerned, short in genome length, such as the case for *P. ubique* and *Prochlorococcus*, is not necessarily disadvantageous as long as it is compensated in time by a capacity information rate. Chromosome length seems to be spacially and physically limited, but replication time lasts as long as life is permitted in the universe.

Our communication model for DNA replication inevitably implies that the principle of DNA replication is information rate optimization. This idea also gives rise to a logical explanation to the problem of junk DNA present in many species genomes. Take the Human genome for example, which is known to contain about 97% junk DNA. A leading conventional explanation surmises junk DNA to be evolutionary left-over, a notion inconsistent with evolutionary optimization. However, if life is to replicate information, then junk or not makes little difference as far as information is concerned — every base carries the same amount of information which is context and observer independent. We also see this in the generalized Chargaff law which both Human and *P. ubique* genomes satisfy. *P. ubique* has adopted a lean and mean genome of 1,308,759 base pairs to build a complete set of biosynthesis pathways. Each replication replicates the machinery only. It is an exception to junk DNA but not to information replication. It is a bargain in itself because of its numerical abundance so that the net information replicated is huge. At the other extreme, we use only 3% of our genome for our replication machinery. It too is a bargain because the partition results in a 32:1 ($= 97$:3) payoff-to-cost gain.

In conclusion, our model suggests that what is worth replicating has to be optimal — be it the best mean rate for the choice of the number of bases; be it the best per-base entropy for non-replicating virus genomes; and be it the best information replication rate for cellular organisms. Given the abstractness of the concept of information and the near impossibility of simulating evolution in laboratories, the most we can hope for is to build empirical consistency for the theory. Expanded data surveys and new experiments specifically designed to determine base pairing and cell replicating times are certainly needed to further test the model.

Acknowledgement: The author gratefully acknowledges the generosity of Dr. Daniel Smith, Department of Biology, Oregon State University, who provided the base frequencies of *P. ubique*, and Dr. David Ussery, Center for Biological Sequence Analysis, Technical University of Denmark, who provided the base frequencies of all other microbes and viruses.

References

[1] Akman L, *et al.*, Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia, *Nat. Genet.*, **32**(2002), 402–407.

[2] Blattner, F.R., The complete genome sequence of Escherichia coli K-12, *Science*, **277**(1997), 1453–1474.

[3] Chisholm, S.W., R.J. Olson, E.R. Zettler, R. Goericke, J.B. Waterbury & N.A. Welschmeyer, A novel free-living prochlorophyte abundant in the oceanic euphotic zone *Nature*, **334**(1988), pp.340–343.

[4] Deng, B., Why is the number of DNA bases 4? *Bul. Math. Bio.*, to appear (2006)

[5] Giovannoni, S.J. *et al.*, Genome streamlining in a cosmopolitan oceanic bacterium, *Science*, **309**(2005), 1242–1245.

[6] Perna, N.T., *et al.*, Genome sequence of the Escherichia coli O157:H7, *Nature*, **409**(2001), 529–533.

[7] Gabrielle Rocap, G. *et al.*, Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation, *Nature*, **424**(2003), pp.1042–1047.

[8] Sergienko, I.V., A.M. Gupal, A.A. Vagis, Complementary principles in encoding bases by one chain in DNA chromosomes, *J. Auto. & Info. Sci.*, **4**(2005), 153-157.

[9] Shannon, C.E., A methematical theory of communication, *Bell System Technical Journal*, **27**(1948), 379–423, 623–656.

[10] Smith, Daniel, private communication, Nov. 2005.

[11] Ussery, D.W. and P.F. Hallin, Genome update: AT content in sequenced prokaryotic genomes, *Microbiology*, **150**(2004), 749–752.

[12] Venter, J.C., *et al.*, The sequence of the human genome, *Science*, **291**(2001), 1304–1351.

Appendix. Recall the definitions for $H_n, T_n(\tau), H(p), T(p,\tau), R(p,\tau)$ from the main text.

For the mean rate R_n of the replication model, we have

$$T_n(\tau) = \frac{2}{n}\sum_{k=1}^{n/2}[\tau_{_{AT}} + \Delta\tau(k-1)] = \tau_{_{AT}}[1 + (\alpha-1)(n-2)/4],$$

using the identity that $1+2+3+\cdots+n = n(n+1)/2$. Hence, the mean rate R_n as presented in the Introduction. Here, $\Delta\tau$ represents a constant pairing time increment so that $\tau_{_{GC}} = \tau_{_{AT}} + \Delta\tau$, under the assumption that the GC pair takes longer to pair than the AT pair does for having an additional hydrogen bond. As for a third, and additional base pairs, it is assumed that $\tau_{56} = \tau_{_{AT}} + \Delta\tau \times 2$, and $\tau_{(2k-1)(2k)} = \tau_{_{AT}} + \Delta\tau \times (k-1)$, etc. Replacing $\Delta\tau$ by $\Delta\tau = \tau_{_{GC}} - \tau_{_{AT}} = \tau_{_{AT}}(\alpha-1)$ gives rise to the formula for $T_n(\tau)$.

For the replication capacity $K(p)$ of the model, we have the following result.

Theorem 1. *The source transmission rate $R(p,\tau)$ has a unique constraint maximum $K(\tau)$ with respect to p when p_i^{1/τ_i} is a constant for all i. In particular, $p_i = p_1^{\tau_i/\tau_1}, \sum_{i=1}^n p_1^{\tau_i/\tau_1} = 1$, and $K(\tau) = -\log_2 p_1/\tau_1 = -\log_2 p_i/\tau_i$.*

Proof. We use the Lagrange method to maximize $R(p,\tau)$ subject to the constraint $g(p) = \sum_{k=1}^n p_k = 1$. This is to solve the joint equations: $\nabla R(p,\tau) = \lambda\nabla g(p), g(p) = 1$, where ∇ is the gradient operator with respect to p and λ is the Lagrange multiplier. Denote $R_{p_k} = \partial R/\partial p_k$, then the first system of equations becomes $R_{p_k} = [H_{p_k}T - H\tau_{p_k}]/T^2 = \lambda g_{p_k} = \lambda$, componentwise. Write out the partial derivatives of H and T and simplify, we have $-(\log_2 p_k + 1/\ln 2)T - H\tau_k = \lambda T^2$ for $k = 1, 2, \ldots, n$. Subtract equation ($k=1$) from each of the remaining $n-1$ equations to eliminate the multiplier λ and to get a set of $n-1$ new equations: $-(\log_2 p_k - \log_2 p_1)T - H(\tau_k - \tau_1) = 0$ which solves to $\log_2 \frac{p_k}{p_1} = R(\tau_1 - \tau_k)$ and hence $p_k = \mu^{\tau_1-\tau_k}p_1$ for all k where $\mu = 2^R = 2^{H/T}$ or $H = T\log_2\mu$. Next we express the entropy H in terms of μ and p_1, τ_1:

$$\begin{aligned} H &= -\sum_{k=1}^n p_k \log_2 p_k = -\sum_{k=1}^n p_k[(\tau_1-\tau_k)\log_2\mu + \log_2 p_1] \\ &= -[\tau_1\log_2\mu - \sum_{k=1}^n p_k\tau_k\log_2\mu + \log_2 p_1] \\ &= -[\tau_1\log_2\mu + \log_2 p_1] + T\log_2\mu, \end{aligned}$$

where we have used $\sum_{k=1}^n p_k = 1$ and $T = \sum_{k=1}^n p_k\tau_k$. Since we have by definition $H = T\log_2\mu$, equating the 2 expressions gives rise to $\log_2 p_1 + \tau_1\log_2\mu = 0$ and consequently $2^R = \mu = p_1^{-1/\tau_1}$ and $p_k = \mu^{\tau_1-\tau_k}p_1 = p_1^{\tau_k/\tau_1}$. Last solve the equation $f(p_1) = g(p) = \sum_{k=1}^n p_1^{\tau_k/\tau_1} = 1$ for p_1. Since $f(p_1)$ is strictly increasing in p_1 and $f(0) = 0 < 1$ and $f(1) = n > 1$, there is a unique solution $p_1 \in (0,1)$ so that $f(p_1) = 1$. The channel capacity $K(\tau) = R(p,\tau) = -\log_2 p_1/\tau_1 = -\log_2 p_k/\tau_k$ follows. This completes the proof. □