Bulletin of Mathematical Biology (2006)
DOI 10.1007/s11538-005-9019-y

ORIGINAL PAPER

Why is the Number of DNA Bases 4?

Bo Deng

Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

Received: 22 November 2004 / Accepted: 14 March 2005

Abstract In this paper we construct a mathematical model for DNA replication based on Shannon's mathematical theory for communication. We treat DNA replication as a communication channel. We show that the mean replication rate is maximal with four nucleotide bases under the primary assumption that the pairing time of the G–C bases is between 1.65 and 3 times the pairing time of the A–T bases.

Keywords DNA replication · Base pairing time · Communication channel · Data rate

1. Introduction

In Claude E. Shannon's mathematical model for communication (Shannon, 1948), there is an information source which produces messages in sequences of alphabets, a transmitter which operates on the messages to produce signals suitable for transmission, a channel through which the signals are sent to a receiver, which usually performs the inverse operation to reconstruct the messages from the signals. This paper proposes a communication model for DNA replication. The proposed conceptual model treats DNA replication as signal transmission. Specifically, the genome of an organism is thought as a message sequence coded in the nucleotides bases, adenine (A), thymine (T), guanine (G) and cytosine (C), and the instantaneous event at which a base is replicated along single strands of the mother DNA sequence is considered the moment when the signal symbol, which is thought to represent the base, is transmitted.

Among many great contributions Shannon made by his theory of communication, two are especially relevant to the model. The first is the separation of the sematic content of a message from the dynamical channel that transmits the message. A particular information source may be, for example, in the context of the Internet, a piece of audio file, or a video clip, or a junk email. In the

E-mail address: bdeng@math.unl.edu.

context of DNA replication, an information source can be the genome of a bacterium, a person, or any organism, each has its own characteristic frequency distribution in the encoding bases. However, the design of a channel is not for a particular message or a particular type of messages. The machinery is for *all* possible messages, regardless of their sematic meanings. How well this machinery does is not judged by its performance on a particular type of source. Instead, it is measured by how well it works for all the possible sources and on average.

The other great contribution of Shannon's that is absolutely vital to all communication systems, our model included, is the quantitative measure of a communication channel in terms of the mean information rate in bits per unit time. In fact, the only intrinsic commodity we buy when signing up an internet carrier, for example, is this mean rate. The means by which the rate is delivered is not essential. It can be carried by DSL or cable or optic fibers, in analog form or digital form, and so on. That is, the mean rate is an intrinsic measure by which distinct systems can be compared.

The data rate for a communication system is defined as

$$R = \lim_{T_c \to \infty} \frac{\log_2 N(T_c)}{T_c},$$

where T_c is the duration (intermittent or continuous) over which the channel is on, $N(T_c)$ is the total *possible* (not actual) number of discrete symbols that can be allowed to go through the channel.

Let $S(T_c)$ be the total number of discrete symbols that are *actually* transmitted over the transmission time T_c, then

$$R = \lim_{T_c \to \infty} \frac{\log_2 N(T_c)/S(T_c)}{T_c/S(T_c)} = \frac{\lim_{T_c \to \infty}[\log_2 N(T_c)/S(T_c)]}{\lim_{T_c \to \infty}[T_c/S(T_c)]} := \frac{H}{\tau},$$

where $H = \lim_{T_c \to \infty} \log_2 N(T_c)/S(T_c)$ is the *entropy* measured in bits per symbol, and the quantity 2^H measures how many possible symbols for each signal symbol transmitted; and $\tau = \lim_{T_c \to \infty} T_c/S(T_c)$ is the average transmission time per transmitted symbol.

To design a communication channel, we first choose the means (telephone lines or optic fibers for example) and the encoding alphabet (states or levels on the electrical waves or optical pulses). Assume there are n bases (states or levels). Then for a message of length S there is a maximal total of $N = n^S$ many possible combinations, if each base is equally probable at every sequence position. Use the exponential base 2 to write $N = 2^{S \log_2 n}$, and we get $H = (\log_2 N)/S = \log_2 n$, the maximal information entropy. However, if the base probabilities are not equally probable, say p_1 for base b_1, p_2 for base b_2, and so on, then the entropy is calculated according to this formula:

$$H(p) := \sum_{k=1}^{n} p_k \log_2(1/p_k),$$

where $\log_2(1/p_k)$ can be interpreted as the information entropy of base k and $H(p)$ the average entropy. To calculate the average base transmission time, let $\tau_1, \tau_2, \ldots, \tau_n$ be the corresponding base transmitting times. Then the average base transmitting time is $\tau(p) := \sum_{k=1}^{n} p_k \tau_k$, measured in time per base. Therefore, the data rate is $R_n(p) = H(p)/\tau(p)$ in bits per time. The important question for the designer of the system to ask is what is the optimal choice in the number of bases n to maximize the data rate $R_n(p)$ with $H(p)$ as large as possible?

2. The mathematical model

We now proceed to translate the conceptual communication model to a mathematical model by incorporating various facts and empirical findings. The goal is to understand the role that four bases play in determining the optimal data rate for the DNA replication. The data rate with optimal $H(p)$ is referred to as the *mean replication rate* or the *mean rate* for short. The full set of the hypotheses is as follows:

1. There is an even number n of nucleotides bases with $n \geq 2$.
2. Replication is done in sequence along a single strand sequence, one base at a time.
3. The bases can be paired as b_{2k-1}, b_{2k} according to the number of their hydrogen bonds, $k+1$.
4. Replication occurs when a base bonds to its complementary base by their hydrogen bonds.
5. The pair-wise hydrogen bonding energies are ordered in the ascending order $E_1 < E_2 < \cdots < E_{n/2}$ for the corresponding base pairs b_{2k-1}, b_{2k} and the bonding energies progress like the natural numbers, i.e., there is an increment ΔE so that

$$E_k = E_1 + \Delta E(k-1), \qquad \text{for } k = 1, 2, \ldots, n/2,$$

   with $E_1 = 9$ kcal/mol and $\Delta E = 3$ kcal/mol.
6. The probability $P_k = P\{b_{2k-1}, b_{2k}\}$ for the b_{2k-1},b_{2k} pair to occur is proportional to its bonding energy E_k, i.e., $P_k/P_i = E_k/E_i$ so that $P_k = E_k/\sum_{i=1}^{n/2} E_i$.
7. Each base occurs equally probable as its complementary base, i.e., the probabilities for base b_{2k-1} and b_{2k} are $p_{2k-1} = p_{2k} = P_k/2$.
8. The pairing times of the paired bases are equal, i.e., $\tau_{2k-1} = \tau_{2k}$, and progress like the natural numbers, i.e., there is an increment $\Delta\tau$ so that

$$\tau_{2k-1,2k} = \tau_{1,2} + \Delta\tau(k-1), \quad \text{for } k = 1, 2, \ldots, n/2.$$

Hypothesis 1 is certainly true for $n = 4$. Biologists have suspected that the current base-4 system started out as a base-2 system with the G–C pairs since they are more stable than the A–T pairs and there are thermally stable amino acids coded by G, C nucleotides alone (Crick, 1968; Reader and Joyce, 2002). So this hypothesis with $n = 2$ is consistent with this belief. Hypothesis 1 with $n \geq 6$ is quite reasonable since in addition to nucleotides A, T, G, C, nucleotides uracil (U) and inosine (I) are also found in RNA (see also Morales and Kool, 1998). Hypothesis 2 is based

on the fact that replication takes place when the complementary nucleotides are bonded to the single strand template in the direction from the sugar base's 5′ carbon to its 3′ carbon. Hypotheses 3 and 4 follow Watson–Crick's complementary base pairing principle (Watson and Crick, 1953) the fact that the A–T, G–C pairs are bonded by hydrogen bonds. The A–T pair has two hydrogen bonds: One hydrogen to oxygen bond N–H···O and one hydrogen to nitrogen bond N–H···N. The G–C pair has three hydrogen bonds: Two hydrogen to oxygen bonds N–H···O and one hydrogen to nitrogen bond N–H···N straddled between. As for Hypothesis 5, the bonding energy of the H···N bond is ∼6 kcal/mol and that of the H···O bond is ∼3 kcal/mol, see Willians and Fraústo da Silva (1996). Hence, $E_1 = 3 + 6 = 9$ kcal/mol and $\Delta E = 3$ kcal/mol. This hypothesis simply assumes that the same pattern persists, i.e., a weaker N–H···O bond is added to the next pair.

Hypothesis 6 requires a greater elaboration. According to Shannon, the 0th-order approximation of a source is the equiprobable distribution, $p_k = 1/n$, so that the entropy $H(p) = \log_2 n$ is the maximum. In fact, this is assumed for any generic communication system, such as the Internet for example, which is designed for all possible sources. Nevertheless, further approximation of all the sources is possible, at least in theory. For English, the 1st-order approximation according to Shannon is the frequency distribution of the alphabet that can only be approximated even if we knew *all* written English since its first written letter because the approximation changes with time as well. Shannon's 1st-order approximation of the genetic code would be much too crude since there are insurmountable limits to what we can know about the genomes of all extinct species. It is a basic fact that the thermal stability (the denature temperature) of a base pair is directly proportional to its hydrogen bond energy. This hypothesis thus postulates that in a purely information source state, like written English, free from the dynamical process of replication, likewise transmission, the DNA source would prefer static states conforming to their thermal stability. In other words, the hypothesized frequency distribution can be considered as a 1st-order approximation of the DNA code. We need to note that this does not mean a particular genome must satisfy nor prefer this 1st-order average.

Hypothesis 7 is a postulation mostly out of convenience for lack of theoretical and empirical arguments either against it or for it. It is a reasonable assumption from the point of view of optimization that equiprobability maximizes information source diversity. The first part of Hypothesis 8 extends Watson–Crick's base pairing principle symmetrically to base pairing times: The time A takes to bond T is the same as T to A, and similarly for the G, C bases. The second part seems reasonable in light of Hypothesis 3 in that the more bonds to pair the longer the pairing times. An empirical finding that RNA transcription is slower in G–C rich regions and faster elsewhere (Uptain et al., 1997) may be viewed as an indirect and qualitative evidence. More discussion is given below on whether or not the natural number progression assumption is a consequence of Hypothesis 3. What is needed at this point is to introduce the leading pairs pairing time ratio

$$\alpha := \frac{\tau_{3,4}}{\tau_{1,2}} = 1 + \frac{\Delta\tau}{\tau_{1,2}},$$

which is to be used as a parameter to illustrate the main result of this model.

We are now ready to calculate the mean rate of this model. Under Hypotheses 6 and 7, the information entropy is

$$H(p) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = \sum_{k=1}^{n/2} P_k \log_2 \frac{2}{P_k}.$$

The average base pairing time is

$$\tau(p) = \sum_{i=1}^{n} p_i \tau_i = \sum_{k=1}^{n/2} P_k[\tau_1 + \Delta\tau(k-1)].$$

For the mean rate, $R_n(p) = H(p)/\tau(p)$, a few simple simplifications are first in order. Factoring out E_1 and cancelling it from both the numerator and denominator of $P_k = E_k/\sum E_i$ shows that P_k and thus $R_n(p)$ depend on the dimensionless parameter $\Delta E/E_1$ as far as parameters E_is are concerned. Divide the equation $\tau(p)$ by τ_1. Then the dimensionless ratio $\tau(p)/\tau_1$ depends on n, $\Delta E/E_1$, $\Delta\tau/\tau_1 = \alpha - 1$ instead. Hence, the rate can be written as $R_n(p) = f(n, \Delta E/E_1, \alpha)/\tau_1$ with f the normalized capacity $\tau_1 R_n(p)$ which is a function of n, $\Delta E/E_1, \alpha$. At the fixed value $\Delta E/E_1 = 1/3$ from Hypothesis 5, $\tau_1 R_n(p)$ is in fact a function of only n and α, the leading base pairs pairing time ratio. Now for $\Delta E/E_1 = 1/3$, Fig. 1 shows the normalized rate $\tau_1 R_n(p)$ as a function of α for various base numbers n. It shows

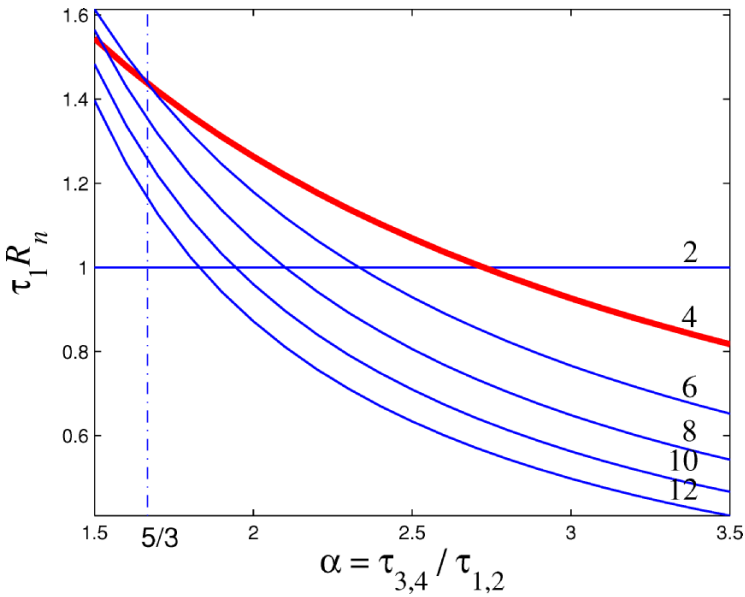

Fig. 1 Numbers next to the curves correspond to the base number n.

that for $1.65 < \alpha < 2.7$, having four bases maximizes the mean rate: $R_4 > R_n$ for all the $n \neq 4$ shown.

3. Discussion

The main question that remains is what is the parameter value α for DNA replication? In other words, what are the base pairing times for the A–T and G–C base pairs? It seems that this question was not encountered in the literature except indirectly (Patel, 2001). We approach this problem by proposing a set of plausible assumptions below.

1. The pairing time t_p is inversely proportional to the bonding energy: $t_p \sim \hbar/E$, with E the bonding energy, $\hbar$ the Planck constant.
2. The DNA backbone bonding time is negligible.
3. The base pairing is done in series: one hydrogen bond a time.

Assumption 1 may follow reasons from quantum mechanics, in particular the Heisenberg uncertainty principle relating energy and time, (Patel, 2001) that $t_p E \sim \hbar$. Assuming this, then the pairing time for the H$\cdots$O bond takes twice as long to complete as the H$\cdots$N bond does, which in turn takes much longer than the covalent bonds along the DNA backbone since covalent bond energies are typically about 20 times greater than hydrogen bond energies. Thus, Assumption 2 follows from Assumption 1. Last by Assumption 3, the leading base pairing times are $\tau_{A,T} \sim \hbar/6 + \hbar/3 = \hbar/2$, $\tau_{G,C} \sim \tau_{A,T} + \hbar/3 \sim 5\hbar/6$, neglecting the covalent bonding times, and the corresponding ratio $\alpha = \tau_{G,C}/\tau_{A,T} \sim 5/3 = 1.667 \in (1.65, 2.7)$, falling right into the R_4 optimal range, and in fact near R_4's absolute maximum. In another alternative scenario, if the stronger H$\cdots$N bond's faster pairing time can be neglected, then $\alpha = 2$ falls safely inside the R_4 optimal range. If the 1st-order approximation of the DNA source by Hypothesis 6 is replaced by the 0th-order approximation by equiprobability distribution, then the optimal interval shifts to $1.825 < \alpha < 3$, which will miss the critical value $\alpha = 1.667$ but still contain $\alpha = 2$.

The statistical mean of the base pairing ratio α remains an open problem. Nevertheless, if the proposed model is a good one for DNA replication, then it will give some good explanations to some outstanding questions. The most important of all, it would imply that life on Earth is where it should be in time. This is because information entropy measures how diverse the per unit length genomes in the pool of life is, and the mean rate measures how much of a diversity that can go through the time bottle neck set up by the channel. Look at the mean rate differently: its reciprocal measures the time needed to replicate a unit bit of diversity. At the minimum replication time needed, each bit of diversity moves through time the fastest, leading to the greatest mutation rate and consequently the fastest adaptation rate. It also leads to the greatest consumption (metabolism) rate, thus out competing and wiping out all non-quaternary systems. In particular, it would explain why DNA evolved away from a base-2 system if indeed life started with a protogenic form of base-2 replication in the G, C bases. For this base G–C system, the mean rate is smaller than that of the base A–T system as shown in Fig. 1 because the former is $(\tau_{A,T}/\tau_{G,C})R_2$ and the latter is R_2 with

$\tau_{A,T}/\tau_{G,C} = 5/3$. At $\alpha = 5/3$, the base-4 system is about 40% faster than the base-2 A–T system, and hence the base-4 system is 133% faster than the base-2 G–C system. Assume life on Earth started about 4 billion years ago, then the base-2 G–C system would set the evolutionary clock backward by 2.3 billion years. If the model is right, it would likely explain why RNAs are also coded in four bases. It is believed that there was an 'RNA world' before the 'DNA world' (Lee et al., 1996; Poole et al., 1998). However, just how RNA replicates itself is not very clear. Assume it does in the same way as DNA does but only in reverse—use itself as the mother template to produce a double strand 'RNA' and then separate itself from the daughter strand. Then the mean rate analysis above equally applies, that is, the mean rate reaches the maximum with four bases provided the leading bases pairing time ratio α is in the R_4 optimal range. (This would also give a plausible scenario as to how DNA first appeared.) It would also give a good explanation as to why there are some odd bases littered amongst the standard A, T, G, C bases. It would suggest their existence to be the result of Nature's relentless, memoryless, and so far unsuccessful attempts to better the quaternary system.

Acknowledgements

The author benefitted from his conversations with, and comments from these colleagues: Irakli Loladze of the Department of Mathematics, Etsuko Moriyama of the Beadle Center for Genetics Research, and Hideaki Moriyama, Xiao-Cheng Zhen, both Department of Chemistry, all at University of Nebraska-Lincoln.

References

Crick, F.H.C., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379.
Lee, D.H., Granja, J.R., Martinez, J.A., Severin, K., Ghadri, M.R., 1996. A self-replicating peptide. Nature 382, 525–528.
Morales, J.C., Kool, E.T., 1998. Efficient replication between non-hydrogen-bonded nucleoside shape analogs. Nat. Struct. Biol. 5, 950–954.
Patel, A.D., 2001. Quantum algorithms and the genetic code. Quantum Phys. Abstract, arXiv:quant-ph/0002037, v3, 6 Feb.
Poole, A.M., Jeffares, D.C., Penny, D., 1998. The path from the RNA world. J. Mol. Evol. 46, 1–17.
Reader, J.S., Joyce, G.F., 2002. A ribozyme composed of only two different nucleotides. Nature 420, 841–844.
Shannon, C.E (1948). A methematical theory of communication, Bell Syst. Tech. J. 27, 379–423, 623–656.
Uptain, S.M., Kane, C.M., Chamberlin, M.J., 1997. Basic mechanisms of transcript elongation and its regulation. Annu. Rev. Biochem. 66, 117–172.
Watson, J.D., Crick, F.H.C., 1953. Molecular structure of Nucleic Acids. Nature 171, 737–738.
Willians, R.J.P., Fraústo da Silva, J.J.R., 1996. The Natural Selection of the Chemical Elements. Clarendon Press, Oxford.