# Mismatch Repair Implies Chargaff's PR2 For Nucleus DNA

Bo Deng[1]

**Abstract: Chargaff's second parity rule (PR2) holds empirically for most types of DNA that along single strands of DNA the base contents are equal for complimentary bases, $\mathsf{A} = \mathsf{T}$, $\mathsf{G} = \mathsf{C}$. A Markov chain model is constructed to track the base evolution of any position along template strands of DNA whose replication is equipped with mismatch repair. Under the key assumptions that the mismatch error rates primarily depend on the complementarity and the steric effect of the nucleotides and that the mismatch repairing process itself makes strand recognition errors, the model shows that the steady state probabilities for the base position to take on one of the 4 nucleotide bases are equal for complimentary bases, and that PR2 is the result of the law of large numbers acting on the base's steady state distribution.**

*Key Words:* Chargaff's second parity rule, Watson-Crick's base paring complementarity, DNA replication, mismatch repair, Markov chain, steady state distribution, the law of large numbers.

**Introduction.** Erwin Chargaff discovered in 1950 ([4]) that the amount of DNA base $\mathsf{A}$ is about the same as that of base $\mathsf{T}$ and the same holds for base $\mathsf{G}$ and base $\mathsf{C}$. This parity rule played an important role in the discovery of the structure of DNA by James Watson and Francis Crick in 1953 ([17]). In return, Watson-Crick's double helical model with its base paring rule — $\mathsf{A}$-to-$\mathsf{T}$ and $\mathsf{G}$-to-$\mathsf{C}$ — gives a mechanistic explanation to Chargaff's complementary parity.

Chargaff and colleagues made a surprising discovery in the late 60s ([5, 6]) that the same rule holds even along the complementary, single strands of the DNA molecules of a common soil bacterium *Bacillus subtilis*. This rule is now referred to as *Chargaff's second parity rule* (PR2) in the literature. Their findings have been extended to many other organisms throughout the years, leading to a recent comprehensive test by [13] which shows that PR2 is nearly universal: It holds for four of the five types of double stranded genomes: the archeal nucleus chromosomes, the bacterial nucleus chromosomes, the eukaryotic nucleus chromosomes, and most double stranded viral genomes. But it fails to hold for some organellar genomes (most noticeably animal mitochondria and plastids), single stranded viral genomes and RNA genomes. The order of magnitude for the PR2-compliance is astonishing — for the Human genome, for example, it amounts to about $3 \times 10^9$ base pairs for the compliant nucleus DNA to a pantry $1.7 \times 10^4$ base pairs for the non-compliant mitochondrion DNA.

Even more surprising is the observation that PR2 cannot be explained by the Watson-Crick DNA model; nor by pure chance; nor by chemistry because it is the purines pair $\mathsf{A}, \mathsf{G}$ and, respectively, the pyrimidines pair $\mathsf{T}, \mathsf{C}$ that are structurally similar. It is equally amazing that there seems to be no logical imperative from Darwin's theory of evolution for the immutability and pervasiveness of the rule. Not surprisingly, mechanistic explanations for PR2 have been sought ([8, 16, 10]),

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

including the stem-loop substitution explanation ([9]), the base inversion and transposition explanation ([1]). However, all proposed explanations to date apply to both PR2 genomes and non-PR2 genomes, leaving their validity and effectiveness to questioning. Various mathematical models were also proposed to explain the rule, but all led to a wrong prediction that single strands of any fixed length would eventually evolve to exhibit the exact rule after long enough time, which is contradicted by the length-dependent property of the rule — the shorter the strand the greater deviation from the parity. The last defect was noticed long ago by [11].

We will adopt in this paper a methodological view that all sciences progress in successions of approximation. For a vast majority of genomes, we propose here that the first order of mechanistic approximation to PR2 is the mismatch repair during replication for all nucleus DNA. As it turns out, our result will not apply to mitochondrion DNA for which mismatch repair is not known to exist for any organism ([15, 3]). We will assume that like all processes, mismatch repair itself is not 100% error proof, and that occasionally it mistakes a replicate strand for a template strand. We will then show that PR2 is the result of this regularly scheduled, systematic, and imperfect process. By this idea of progressive approximation, we envision that other less frequent and less systematic processes — such as the stem-loop substitution, the base inversion and transposition — may count for some secondary or more restrictive causes to PR2 genomes without mismatch repairs. We also adopt the view that when other considerations are equal a simpler model can be considered to be the primary approximation to the solution of a problem, which when applied to this case implies that mismatch repair can be considered as the primary cause for nucleus PR2.

**The Mathematical Model.** The main idea for our mathematical model is to model the mismatch repair as a Markov process for each position of the template strands. When combined with the law of large numbers, the Markovian steady state probability distribution will give rise to the length-dependent PR2. The mathematical model of DNA replication with mismatch repair assumes the following.

  **Mismatch Repair Error (MRE) Model:**

  (a) DNA replication makes nucleotide mismatch errors.

  (b) At the moment of base replication, the mismatch error occurs independent of the nucleotide base position on the template single strand of DNA.

  (c) The match probabilities satisfy the following assumptions:

  i. The match probabilities between complementary bases are equal with respect to the complementary pairs:

$$m\{A \to T\} = m\{T \to A\} \ \text{ and } \ m\{G \to C\} = m\{C \to G\}$$

  ii. The match probability of a base to its complementary or itself is the same as that of the base's complementary to the complementary or itself:

$$m\{A \to A, \text{or } T\} = m\{T \to A, \text{or } T\} \ \text{ and } \ m\{G \to G, \text{or } C\} = m\{C \to G, \text{or } C\}$$

2

iii. The conditional mismatch probability of a pyrimidine to its non-complementary purine equals that of its complementary purine to the purine's non-complementary pyrimidine:

$$m\{A \to C | A \not\to T\} = m\{T \to G | T \not\to A\}$$

$$m\{G \to T | G \not\to C\} = m\{C \to A | C \not\to G\}$$

(d) Mismatch repair makes strand recognition errors independent of complementary bases, and as a result it make base substitution errors.

Hypothesis (a) is self-evident, which applies to all DNA, with or without mismatch repair. Hypothesis (b) is a localized and symmetrical assumption that each base position is as critical or ordinary as any other base position. It should be taken to be a *primary* approximation of this aspect of the replication. This means any assumption about global interactions along single strands of DNA, such as the stem-loop hypothesis of [2], or codon position bias asymmetric assumption from [11], can be taken as secondary approximations or corrections for future refinement of the model.

Hypotheses i,ii of (c) imply that the mismatch probability of a base to itself is the same as that of the base's complementary to the complementary:

$$m\{A \to A\} = m\{T \to T\} \text{ and } m\{G \to G\} = m\{C \to C\}$$

Alternatively, Hypothesis ii of (c) can also be stated in terms of the mismatch conditional probabilities:

$$m\{A \to A | A \not\to T\} = m\{T \to T | T \not\to A\}$$

$$m\{G \to G | G \not\to C\} = m\{C \to C | C \not\to G\}$$

In other words, Hypotheses i,ii of (c) are made mainly on the complementarity of bases, which would be automatically true if DNA were a binary code in either the $AT$-pair or the $GC$-pair.

On the other hand, however, Hypothesis iii of (c) is made mainly on the steric characteristics of the pyrimidines and purines, i.e., when an $A \not\leftrightarrow T$ (resp. $G \not\leftrightarrow C$) mismatch occurs, the conditional mismatch probability between $A \to C$ (resp. $G \to T$) mismatch is the same as the $T \to G$ (resp. $C \to A$) mismatch. As a result of Hypotheses i,ii,iii of (c), the remaining conditional mismatch probabilities are forced to satisfy:

$$\begin{aligned} m\{A \to G | A \not\to T\} = m\{T \to C | T \not\to A\} \\ m\{G \to A | G \not\to C\} = m\{C \to T | C \not\to G\} \end{aligned} \tag{1}$$

More details can be found from the match/mismatch probability diagrams of Fig.1. We note that hypotheses similar to (c) can also be found in the literature, c.f. [16, 10]

What separates our model from all others is Hypothesis (d). Although similar assumptions such as base inversion, inverted transposition, stem-loop substitution, codon position-bias were made for other models, the precise and systematic mechanisms that made such operations possible
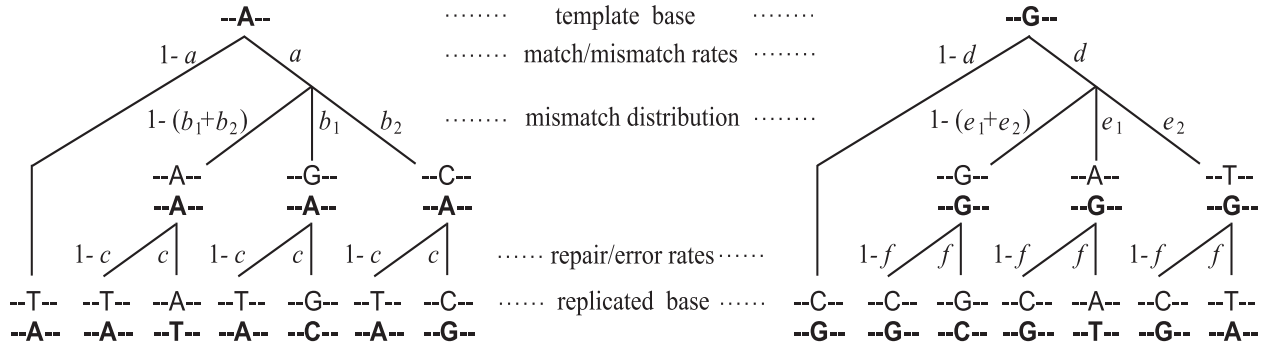
Figure 1: MRE charts for base A and base G. Parameters $a, d$ denote the mismatch probabilities $m\{A \not\to T\}, m\{G \not\to C\}$, respectively, and $b_i, e_i$ denote the conditional mismatch probabilities. For example, $b_2 = m\{A \to C | A \not\to T\} = m\{T \to G | T \not\to A\}$ and similarly, $e_2 = m\{G \to T | G \not\to C\} = m\{C \to A | C \not\to G\}$.

were either not known well enough or conjectured too broadly for all types of DNA, to which we know by the result of [13] that PR2 does not always apply. As a result, none is incorporated into the current model, nor is any repairing mechanism outside the phase of replication such as excision repair to spontaneous deamination of cytosine. In contrast, it appears that most types of double-stranded chromosomes, nucleus chromosomes in particular, are equipped with mismatch repair ([14]). More interestingly, mismatch repair in mitochondria of any organism is not known to exist according to [15], even though other types of repair mechanisms may exist ([3]) which do not negate the model under consideration. This distinction between DNA having or not having mismatch repair is consistent with the applicability of PR2 established in [13]. As a result of this hypothesis, our MRE model applies only to DNA types with mismatch repair satisfying these four hypotheses.

**The Result.** From a modeling perspective, Hypotheses (c,d) imply that the MRE model is a Markov process model for any arbitrary position of any single strand of DNA since the transition probability at the position depends only on its current base. The Markov model is illustrated in Fig.1, referred to as the *mismatch repair error* chart, or MRE chart for short. Take, for example, a nucleotide A on a template single strand of DNA as shown at the top of the left chart. With a probability $0 < 1 - a < 1$, the process correctly replicates A's complementary base T, showing down the left most branch, but incorrectly with $0 < a < 1$ probability. Of that fraction of mismatches, for a fraction of $0 < b_1, b_2$ each the process mismatches it with a G or a C, respectively, with $b_1 + b_2 < 1$. For the remaining $1 - (b_1 + b_2)$ fraction of mismatches, an A is mismatched to the original A. These assumptions are represented by the middle three branches.

For organisms which proof read and repair mismatches, the MRE chart continues one level down to the bottom branches. In such cases, Hypothesis (d) assumes that they do not always distinguish the replicative strand from the template strand 100% of the time, and make strand recognition errors independent of the replicating bases in question. Thus, for $0 < c < 1$ fraction of time, however insignificant or small it may be at this point, the process mistakes the replicative

4

strand for the replicating template, and proceeds to turn the original base A into a base T if the mismatched pair is an A-A pair, with the hyphenated second base position being the template base. Similarly, with a probability $0 < c < 1$ each, a G-A pair is transformed to a G-C pair and a C-A pair to a C-G pair, all changing the template base A to a non-complementary base. With $1 - c$ probability, the process correctly identifies the template strand, and proceeds to keep the original base A in the template strand when the replication is completed. This particular illustration shows the case that the original mismatch error is corrected by a base T, which needs not to be the case. In fact, what is only assumed and important is that with the $1 - c$ probability the original base A is *preserved* along the template strand upon replication. The replicated new base along the complementary strand may or may not be the complementary base of the old base upon the completion of the mismatch repair. That is, the most general depiction of the chart would replace the base T by a place holder, O, for any of the four bases.

Interchanging A and T and interchanging the parameters $b_1, b_2$ in base A's MRE chart gives rise to base T's MRE chart if the replicating base at the top starts with a T. Because of Hypotheses (c,d), all the probability distributions for T's chart are the same as base A's chart except for the conditional mismatch probabilities of Hypothesis iii(c) and the consequence Eq.(1) of Hypothesis (c), which are accounted for by interchanging $b_1$ and $b_2$. The transition probabilities from base T are listed on the second row of the transition probability matrix $P$ below.

Similarly, the diagram on the right is base G's MRE chart with probability distributions in $d, e_1, e_2, f$, which may not be respectively the same as $a, b_1, b_2, c$ for A and T. In particular, $a$ and $d$ are very likely not to be the same, so are for $b_i$ and $e_i$. According to Hypothesis (d), we have $c = f$. But there will be no changes to the result if we keep $c, f$ distinct. Also, interchanging G and C in G's MRE chart gives rise to C's MRE chart.

The mismatch repair charts allow us to derive the transition probabilities between bases one replicating generation a time. More specifically, take the A's MRE chart for an example. Let $p\{A \rightarrow A\}$ be the probability that a base A from a template strand remains an A upon replication. Similarly, let $p\{A \rightarrow T\}$ be the the probability that a base A from the template strand is substituted by a T upon replication, and similar notation applies to $p\{A \rightarrow G\}, p\{A \rightarrow C\}$. Then these probabilities can be tabulated from probability distributions from the MRE chart as follows:

$$p\{A \rightarrow T\} = a(1 - (b_1 + b_2))c$$
$$p\{A \rightarrow G\} = ab_1c$$
$$p\{A \rightarrow C\} = ab_2c$$
$$p\{A \rightarrow A\} = 1 - p\{A \rightarrow T\} - p\{A \rightarrow G\} - p\{A \rightarrow C\} = 1 - ac$$

The first expression, for example, is obtained by following from the top of the chart down the direct branches leading to the substituting base T and multiplying all the probabilities along the branches. The same for the transition probabilities from A to G and C respectively. The probability from A to A can either be obtained by the formula above or by summing all probabilities of the A-to-A branches (four branches in all) of the chart. Exactly the same tabulation gives the transition

probabilities for all other bases.

Using matrix entry notation, we denote

$$p_{11} = p\{\mathsf{A} \to \mathsf{A}\},\ p_{12} = p\{\mathsf{A} \to \mathsf{T}\},\ p_{13} = p\{\mathsf{A} \to \mathsf{G}\},\ p_{14} = p\{\mathsf{A} \to \mathsf{C}\}.$$

That is, the row and column numbers, 1, 2, 3, 4, are in correspondence with $\mathsf{A}$, $\mathsf{T}$, $\mathsf{G}$, $\mathsf{C}$, respectively. Similar notation extends to other bases as well. As a result we obtain the following transition probability matrix for our MRE model:

$$P = [p_{ij}] = \begin{bmatrix} 1 - ac & a(1 - (b_1 + b_2))c & ab_1c & ab_2c \\ a(1 - (b_1 + b_2))c & 1 - ac & ab_2c & ab_1c \\ de_1f & de_2f & 1 - df & d(1 - (e_1 + e_2))f \\ de_2f & de_1f & d(1 - (e_1 + e_2))f & 1 - df \end{bmatrix}$$

What follows are standard textbook properties of transition probability matrixes of Markov chains (c.f.[12]).

1. Denote the $n$th iterate of the transition matrix by

$$P^n = [p_{ij}^{(n)}].$$

   Then we know that it is again a transition probability matrix with its entry $p_{ij}^{(n)}$ representing the probability of a base $i$ becoming a base $j$ after the $n$th generation of replication.

2. Because the transition matrix $P$ has all positive entries (hence is irreducible and ergodic), by the Perron-Frobenius Theorem, the limit $\lim_{n \to \infty} p_{ij}^{(n)}$ exists and the limit is independent of the initial base $i$: $\lim_{n \to \infty} p_{ij}^{(n)} = w_j$, satisfying $0 < w_j < 1$ and

$$w_1 + w_2 + w_3 + w_4 = 1.$$

   In terms of DNA replication with mismatch repair, probability $w_j$ is the *steady state probability* of finding base $j$ at any base position of any single strand of DNA. This means, regardless of the initial base $i$ at that position, after sufficiently many generations of replication, the probability of finding base $j$ at the position is $w_j$. We denote the four steady state probabilities by $w_1 = p\{\mathsf{A}\}, w_2 = p\{\mathsf{T}\}, w_3 = p\{\mathsf{G}\}, w_4 = p\{\mathsf{C}\}$.

3. The transition matrix $P$ has $\lambda_1 = 1$ to be the largest eigenvalue in magnitude and it is simple. Moreover, the steady state probability vector $w = [w_1, w_2, w_3, w_4]$ is the only left eigenvector not counting scalar multiple (or the transpose $w^t$ is the only right eigenvector of eigenvalue 1 for the transpose matrix $P^t$). Because of this property, we can use the eigenvector equation $wP = w$ to find $w$ explicitly as

$$w = \frac{1}{2(a(b_1 + b_2)c + d(e_1 + e_2)f)}[d(e_1+e_2)f,\ d(e_1+e_2)f,\ a(b_1+b_2)c,\ a(b_1+b_2)c],$$

   with equal probabilities $w_1 = w_2$, $w_3 = w_4$ for complementary bases. This, to be explained later, is the basis for the empirical law of PR2.
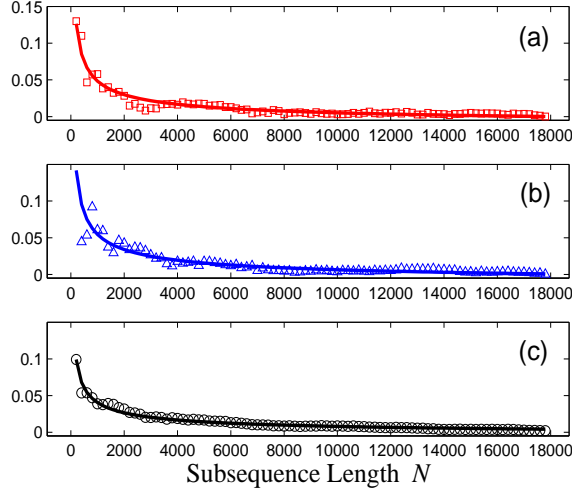
Figure 2: Homo sapiens chromosome 22 genomic contig ref|NT_113818.1|Hs22_111537:1-17927 (from Human Genome Project) with length $L = 17,927$ and ensemble frequency $F_L = (0.1993, 0.2007, 0.2917, 0.3083)$. (a) In the $5' \rightarrow 3'$ direction, the sequence of ensemble frequencies $F_N$ with length $N = 1, 2, \ldots, L$ is generated. The convergence error sequence, $\sum_{i=A,T,G,C} |F_N(i) - F_L(i)|$, is plotted against $N$ together with its best fit to the curve $a + b/\sqrt{N}$. (b) The same plot except for the opposite $3' \rightarrow 5'$ direction. (c) The same plot except (1) the data is averaged over 10 runs and (2) each run is done for a random permutation of the original sequence.

4. By the Perron-Frobenius Theorem, the remaining eigenvalues of $P$ are less than 1 in magnitude. In this case, they can be explicitly found as:

$$\lambda_2 = 1 - a(b_1 + b_2)c - d(e_1 + e_2)f$$
$$\lambda_{3,4} = 1 - ac - df + \tfrac{1}{2}(a(b_1 + b_2)c + d(e_1 + e_2)f) \pm \tfrac{1}{2}\sqrt{D}, \quad \text{where}$$
$$D = (-2ac + 2df + a(b_1 + b_2)c - d(e_1 + e_2)f)^2$$
$$+ 4a(b_1 - b_2)cd(e_1 - e_2)f$$

It has the property that for any initial probability distribution $q = [q_1, q_2, q_3, q_4]$ with $q_1 + q_2 + q_3 + q_4 = 1$, $qP^n$ converges to $w$ at a rate no greater than the order of $|\mu|^n$ with $\mu = \max_{2 \leq i \leq 4}\{|\lambda_i|\} < 1$. This gives a temporal estimate for the convergence rate to the steady state probability at the $n$th generation of replication.

We now conclude that Hypothesis (b), Property 2, and the law of large numbers imply the empirical PR2. More specifically, let $F_L\{i\}$ be the length-averaged frequency of base $i$, called the *ensemble frequency*, then we must have $\lim_{L \rightarrow \infty} F_L\{i\} = p\{i\}$ for $i = A, T, G, C$. To see this, we borrow a prototypical explanation from coin tossing: The steady state probabilities of the head and tail of a coin can be approximated by tossing an ensemble of $L$ many identical coins and counting the ensemble frequencies in which case the greater the assemble size $L$ is the closer the ensemble frequency to the steady state probability distribution becomes. In fact, according to the law of large numbers (or the Central Limit Theorem), the convergence rate has the order of $1/\sqrt{L}$. Fig.2

7

gives a typical simulation of the convergence for a short contig of the Human chromosome 22. Simulations with longer sequences are not shown here because in general the longer a sequence the better the convergence fit becomes.

**Discussion.** It is interesting to note a contrast between mismatch repair and other possible or conjectured processes such as base inversion, transposition, stem-loop substitution, codon position-bias asymmetry. For organisms without mismatch repair, the MRE chart stops at the first or the second branches with the template base well-entrenched in its position. It is also important to note that the model starts with a single stranded DNA molecule during replication and ends with the completion of replication. This means that any damage or repair that may take place before or after replication is not a part of the model at its current formulation. And if such types of damages and repairs are not as systematic and frequent as the mismatch repair is, their effects may well be some secondary corrections to the PR2. We also note that the PR2 regularity or the lack of it for genomes *without* mismatch repair is neither a confirmation nor a contradiction to our model. In other words, the underlining assumption in mismatch repair is a sufficient but not a necessary condition for PR2. Processes other than mismatch repair may result in PR2 as well.

Past modeling attempts were not successful mathematically because all models failed to make the critical distinction between the steady state probabilities at individual base positions and the ensemble frequencies along whole strands of DNA molecules, leading to erroneous fit to the latter. To a lesser extent, some of them failed because their proposed causes apply to both PR2 compliant and non-compliant genomes.

It appears that the main purpose of mismatch repair is to enforce replication fidelity. But one can argue with an equal, perhaps even greater logical soundness that mismatch repair is to enforce Watson-Crick's base complementarity even though it may change a template complementarity by either transposing it or substituting it for a different kind. Thus, if the proposed model is correct, implying the pervasive PR2 for nucleus DNA, then it will raise the question that is PR2 also a genomic fidelity that needs to be preserved? In other words, is PR2 a merely side-effect or does it serve some functional purposes? If it is the latter, what is the evolutionary advantage of having it? An answer to this question may be found in the mathematical model of DNA replication proposed in [7] and it will be addressed elsewhere.

The MRE model is suitable for experimental test. More specifically, an experiment can start with many identical double strands of DNA about 6,000 or more bp long (c.f. Fig.2) with an ensemble frequency disobeying PR2. Divide them into two groups. Replicate one group by a replicator with mismatch repair and replicate the other group by a replicator without mismatch repair. After a *sufficiently* long period of time, calculate the averaged ensemble frequency from each group and compare. The experiment would be supportive of the model if the ensemble frequency of the mismatch-repairing replicator is closer to PR2 than that of the null replicator. The model can also be rejected if one can find a non-PR2 genome with mismatch repair.

# References

[1] Albrecht-Buehler, G., Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions, *PNAS*, **103**(2006), pp.17828–17833.

[2] Bell, S.J. & D.R. Forsdyke, Accounting units in DNA, *J. Theor. Bio.*, **197**(1999), pp.51–61.

[3] Berneburg M., Y. Kamenisch, and J. Krutmann, Repair of mitochondrial DNA in aging and carcinogenesis, *Photochem Photobiol Sci.*, **5**(2006), pp.190–198.

[4] Chargaff, E. Structure and function of nucleic acids as cell constituents, *Fed. Proc.*, **10**(1951), pp.654–659.

[5] Rudner, R., J.D. Karkas, E. Chargaff, Separation of *B. subtilis* DNA into complementary strands, III. Direct Analysis, *Proc. Natl. Acad. Sci.*, **60**(1968), pp.912–922.

[6] Rudner, R., Karkas, J.D., Chargaff, E. Separation of microbial deoxyribonucleic acids into complementary strands, *Proc. Natl. Acad. Sci.*, **63**(1969), pp.152–159.

[7] Deng, B., Why is the number of DNA bases 4? *Bul. Math. Bio.*, **68**(2006), pp.727-733.

[8] Fickett, J., D. Torney, and D. Wolf, Base compositional structure of genomes, *Genomics*, **13**(1992), pp.1056-1064.

[9] Forsdyke, D.R, J.R. Mortimer, Chargaff's legacy, *Gene*, **261**(2000), pp.127–137.

[10] Lobry, J.R. and C. Lobry, Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.*, **16**(1999), pp.719–723.

[11] Lobry, J.R., A nice wrong model for the evolution of DNA base frequencies. *Physica A*, **273**(1999), pp.100–103.

[12] Meyn, S.P. and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, 1993.

[13] Mitchell, D., R. Bridge, A test of Chargaff's second rule, *Bioche. Biophy. Res. Comm.* **340**(2006), pp.90–94.

[14] Modrich, P., Strand-specific mismatch repair in mammalian cells, *J. Biol. Chem.*, **272**(1997), pp.24727–24730.

[15] Modrich, P., personal email communication, (2006)

[16] Sueoka, N., Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, *J. Mol. Evol.*, **40**(1995), pp.318–325.

[17] Watson, J.D. and F.H.C. Crick, Molecular structure of nucleic acids, *Nature*, **171**(1953), pp.737–738.

**Supplemental Notes for Review Only.** A recent model by [1] concludes that the convergence rate to PR2 is in the order of $(1/L)^n$ where $L$ is the genome length and $n$ is the number of replication. It leads to the same wrong length-independent PR2 as pointed out in [11]. For a fixed $n$, the length dependent estimate is also way-off. Fig.3 shows a comparison between the case with $n = 1$ against the case with $n = 1/2$ by the law of large numbers. The case of $n > 1$ is much worse than $n = 1$ when all comparing to $n = 1/2$.
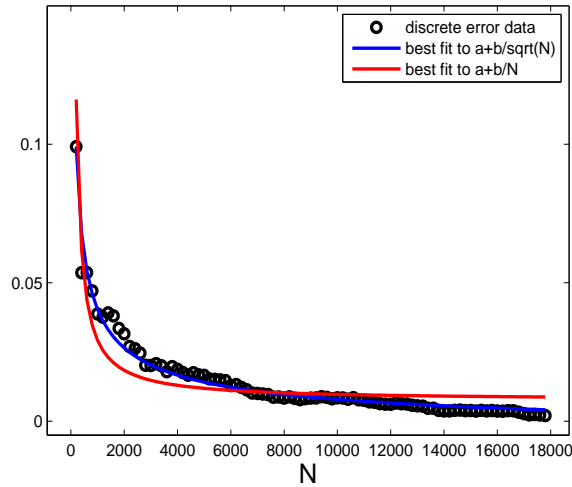


Figure 3: Homo sapiens chromosome 22 genomic contig ref|NT_113818.1|Hs22_111537:1-17927 (from Human Genome Project) with length $L = 17,927$ and ensemble frequency $F_L = (0.1993, 0.2007, 0.2917, 0.3083)$. The same plot as Fig.2(c) except that contrasting fit to $a + b/N$ is included.

Excerpt from Dr. Modrich's reply to an inquiry of the author (Email, received at 1:34pm, Nov 21, 2006, [15]):

> "Although yeast has a nuclear-encoded mitochondrial-targeted MutS homolog (MSH1), the disruption of which leads to mitochondrial degeneration, a mammalian homolog of MSH1 has not been identified, and to my knowledge no one has convincingly demonstrated replication error correction by mismatch repair in mitochondria of any organism. We spent some time looking for such a reaction unsuccessfully–but then these are all negative results."