Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 297–305
GENOMICS
www.elsevier.com/locate/ygeno

Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences

Guenter Albrecht-Buehler *

Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

Received 21 March 2007; accepted 21 May 2007
Available online 20 June 2007

Abstract

In genome duplexes that exceed 100 kb the frequency distributions of their trinucleotides (triplet profiles) are the same in both strands. This remarkable symmetry, sometimes called Chargaff's second parity rule, is not the result of base pairing, but can be explained as the result of countless inversions and inverted transpositions that occurred throughout evolution (G. Albrecht-Buehler, 2006, Proc. Natl. Acad. Sci. USA 103, 17828–17833). Furthermore, comparing the triplet profiles of genomes from a large number of different taxa and species revealed that they were not only strand-symmetrical, but even surprisingly similar to one another (majority profile; G. Albrecht-Buehler, 2007, Genomics 89, 596–601). The present article proposes that the same inversion/transposition mechanism(s) that created the strand symmetry may also explain the existence of the majority profile. Thus they may be key factors in the creation of an almost universal "format" in which genome sequences are written. One may speculate that this universality of genome format may facilitate horizontal gene transfer and, thus, accelerate evolution.

Keywords: Trinucleotide distribution; Transpositions; RNA world; CG-to-TT conversion

Introduction

The existence of a majority triplet profile of the natural genomes

The vast amount of sequence data available today has left no doubt that every species has its own, species-specific coding sequences. Naturally, one may therefore assume that the frequency distributions of their codons are also different from species to species. Yet, as was shown recently [2], that is not the case. Counting not only the codons of genes, but all trinucleotides (triplets $\triangle$) of the coding and noncoding regions in a large variety of genomes, there were essentially only three classes of such triplet distributions among all organisms and DNA containing organelles. One class, called the "majority class," contains the genomes of most organisms ranging from *Rickettsia* to primates. Their frequency distribution $f_M(\triangle)$, which was called the "majority triplet profile," or "majority profile" for short, appears to be a surprisingly universal property of genomes regardless of taxa or species.

The majority profile is shown in Fig. 1. The shaded area covers the standard deviation of 31 genomes that belonged to different organisms and ranged from *Rickettsia* to humans, including chimpanzee, mouse, zebrafish, maize, *Streptococcus pneumoniae, Arabidopsis, Xenopus laevis,* yeast, *Bacillus subtilis, Anopheles*, and others.

The universality of strand symmetry

Plotted as a function of the 64 possible triplets $\triangle = \{n_1, n_2, n_3\}$ (n_i=A, C, T, G), the majority profile appears to be a rather complex function. In part, the apparent complexity is a consequence of the particular order of the triplets along the abscissa ("canonical order" [1]). Yet, hidden in this function, and independent of any ordering convention, is the remarkable property that the frequency of each triplet is equal to the frequency of its reverse complement.

Of course, not every function that fits into the shaded area of Fig. 1 fulfills this condition. Nevertheless, the triplet profile of each individual genome that contributed to Fig. 1 fulfilled it

* Fax: +1 312 503 7912.
E-mail address: g-buehler@northwestern.edu.

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.05.010

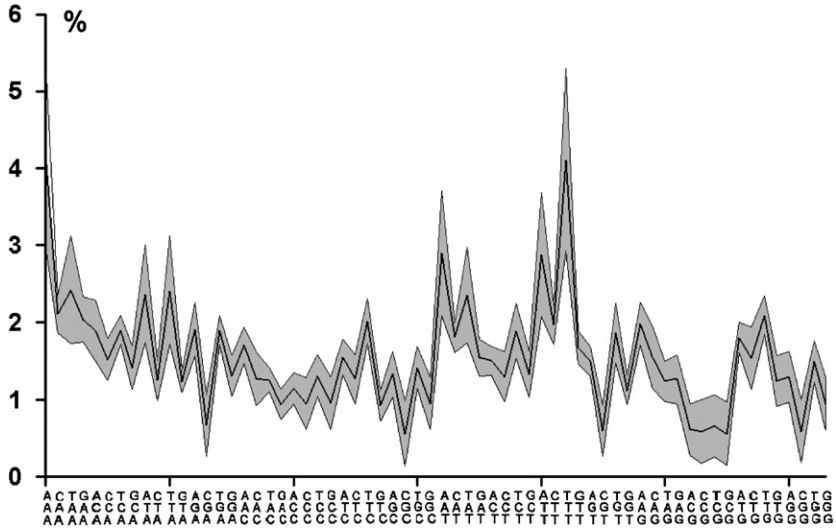

Fig. 1. Majority triplet profile [1,2] (abscissa, triplets to be read from bottom to top; ordinate, fraction of triplets of entire genome). The shaded area covers the range of the standard deviation computed from the triplet profiles of 31 chromosomes of different organisms ranging from *Rickettsia* to humans, including human, chimpanzee, mouse, zebrafish, maize, *Streptococcus pneumoniae, Arabidopsis, Xenopus laevis*, yeast, *Bacillus subtilis, Anopheles*, and others.

quite accurately. In other words, for each of these genomes the triplet profile of the Watson strand was equal to the triplet profile of the Crick strand.

This peculiar strand symmetry has sometimes been called the validity of "Chargaff's second parity rule" [1,3,4]. As shown earlier, it is not a consequence of base pairing, but can be explained as the result of countless inversions and inverted transpositions (henceforth abbreviated as "inversions/transpositions") that the genomes suffered throughout evolution [1].

The effects of large numbers of inversions/transpositions on strand symmetry

The principle effect of such large numbers of inversions/transpositions on strand symmetry is illustrated in Fig. 2. Each duplex DNA is depicted as a pair of straight ribbons labeled as Watson or Crick. The four nucleotides are represented by shades of gray that color the various segments of the ribbons (Fig. 2a). For the sake of simplicity I assumed that all inverted transposons had a constant size (see frames in Figs. 2b and 2c labeled "inv/tp").

The illustration starts with the simplest possible situation of a duplex consisting of a poly(A) strand and its complementary poly(T) strand (Fig. 2b, "0"). At this stage the Watson strand contains only A's, AA's, and AAA's, but no T's, TT's, or TTT's. Likewise, the Crick strand has only T's, TT's, and TTT's, but no A's, AA's, or AAA's. Obviously, there is no symmetry between these strands.

The situation changes after the first inverted transposition has carried some T's to the Watson strand while carrying an equal number of A's to the Crick strand (Fig. 2b, "1", "2"). At this point not only do the complementary nucleotides appear on either strand, they also generate some mixed triplets such as ATT, TTA, AAT, and TAA for the first time on both strands. As the process continues and the number of randomly placed inverted transpositions increases, the distributions of A's, T's, and their corresponding doublets and triplets become increasingly the same. (Please note that the sequences do not become the same, but only their mono-, di-, and trinucleotide distributions do.) A more detailed analysis shows that the equalization of the nucleotide distributions grows exponentially with the number of inversions/transpositions [1].

Similarly, if the initial duplex contains all four nucleotides in some arbitrary ratio, the strands become exponentially more symmetrical with the increasing number of inversions/transpositions. An example is shown in Fig. 2c.

The majority profile as a result of large numbers of inversions/transpositions

By definition of strand symmetry, every such series of inversions/transpositions creates a triplet profile that is the same for both strands. However, this common triplet profile may be different for different starting sequences. The present article will identify the particular starting conditions that yield a triplet profile that will turn out to be identical to the majority profile. In this sense one may consider not only the strand symmetry to be a product of inversions/transpositions, but the majority profile as well.

Results

The stochastic expectation

What triplet profile could one expect if genomes were completely random sequences (stochastic expectation)? Consider the sequence of nucleotides as the result of a random picking from an urn of well-mixed balls labeled as A, C, G, and T. Every selected ball determines the next nucleotide in the sequence and is afterward returned to the urn.

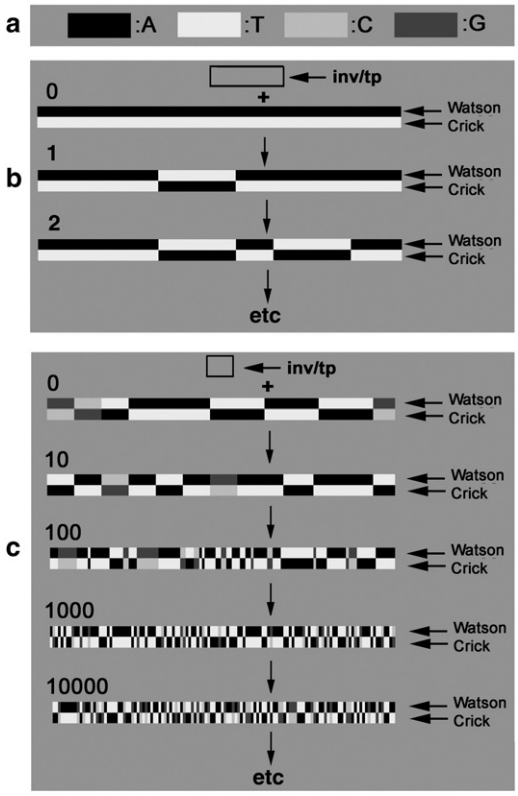

Fig. 2. Illustration of the effects of large numbers of inversions/transpositions on the strand symmetry. Each duplex DNA is depicted as a pair of straight ribbons labeled as "Watson" or "Crick." For the sake of simplicity I assumed that all inverted transposons had a constant size (see frames in (b and c), labeled "inv/tp"). (a) Color coding of the four nucleotides by shades of gray that color the various segments of the ribbons. (b) Equalization of the numbers of A's and T's in the case of a duplex consisting of a poly(A) strand and its complementary poly (T) strand ("0"). Obviously, initially there is no symmetry between these strands. As the number of randomly placed inversions increases the inversions carry increasing numbers of T's to the Watson strand while carrying an equal number of A's to the Crick strand ("1", "2"). They also generate some mixed triplets such as ATT, TTA, AAT, and TAA for the first time on both strands. As the process continues and the number of randomly placed inverted transpositions increases, the distributions of A's, T's, and their corresponding doublets and triplets become increasingly the same. A more detailed analysis shows that the equalization of the nucleotide distributions grows exponentially with the number of inversions/transpositions [1]. (c) Similarly, if the initial duplex contains all four nucleotides in some arbitrary ratio, the strands become exponentially more symmetrical with the increasing number of inversions/transpositions as indicated by the numbers at each duplex.

Assume that the probabilities of the balls are p_0(A), p_0(T), p_0(C), and p_0(G). They must fulfill the normalization requirement

$$p_0(A) + p_0(T) + p_0(C) + p_0(G) = 100\% \quad (1)$$

Then the probability $f_0(\Delta)$ of drawing the triplet $\Delta = \{n_1, n_2, n_3\}$ at random from the urn is given by

$$f_0(\Delta) = p_0(n_1) \cdot p_0(n_2) \cdot p_0(n_3) \quad (2)$$

In the case of the actual genomes of the majority class one can easily calculate from the majority profile their base compositions as p_M(A)=30.6%, p_M(T)=30.6%, p_M(C)=19.4%, p_M(G)=19.4%. It corresponds to an AT content of 61.2% and a CG content of 38.8%. Obviously, these genomes obey Chargaff's second parity rule for mononucleotides, i.e., p_M(A)=p_M(T) and p_M(C)=p_M(G) on each single strand.

In contrast, the present calculation of the stochastic expectation cannot presuppose a similar symmetry a priori, as a previous study had interpreted it as the result of numerous inversions/transpositions that occurred during the evolution of these genomes [1]. Hence, it was assumed (somewhat arbitrarily) that

$$p_0(A) = 20\%, p_0(T) = 36\%, \quad p_0(C) = 25\%, p_0(G) = 19\% \quad (3)$$

The effect of the choice of the particular parameters on the results will be discussed later.

A specially designed computer program created 8-Mb large, random genome sequences with the particular base composition of Eq. (3) and confirmed by direct counting that their triplet profile (Fig. 3b) matched exactly the prediction of Eq. (2).

Such artificial genome sequences represent the stochastic expectation of genomes randomly arising from a large pool of nucleotides as postulated, for example, by the so-called "RNA world" [5–9]. Obviously, there are astronomically large numbers of such genomes that have different sequences, yet nevertheless have the same triplet profile of Eq. (3). However, the genome sequences so generated do not fulfill Chargaff's second parity rules (see Fig. 4b), nor do they belong to the majority class (see Fig. 3b).

If the quality of fitting the profile of randomly generated genomes to the majority profile were the goal of this article, one could improve it considerably by introducing the formalism of the so-called Markov chains. However, the intention of this article is not to fit the majority profile, but to explore the effects of inversions/transpositions on stochastically created, "primordial" polynucleotides. Therefore, a special computer program was designed that subjected stochastic-expectation genomes to a large number of inversions/transpositions after applying various small sequence modifications. Since this article focuses only on the resulting triplet distributions, the actual location of an inverted transposon was irrelevant for the outcome and, therefore, ignored. In other words, the simulations involved inversions only.

Sequence modification

According to our previous results, applying a large number of inversion/transpositions to the stochastic-expectation genomes must generate sequences that fulfill Chargaff's second parity rules [1]. However, this process did not immediately yield triplet profiles that resembled the majority profile. Therefore, it was necessary to modify the stochastic-expectation genomes beforehand in a certain way. I found that the simplest modification yielding a majority profile after inversions/transpositions was to change 60% of CG dinucleotides into TT dinucleotides. This modification was expressed by a "conversion rate" of $\rho = 0.6$.

Even though a conversion rate of 60% may sound like a large modification, the number of sequence alterations was

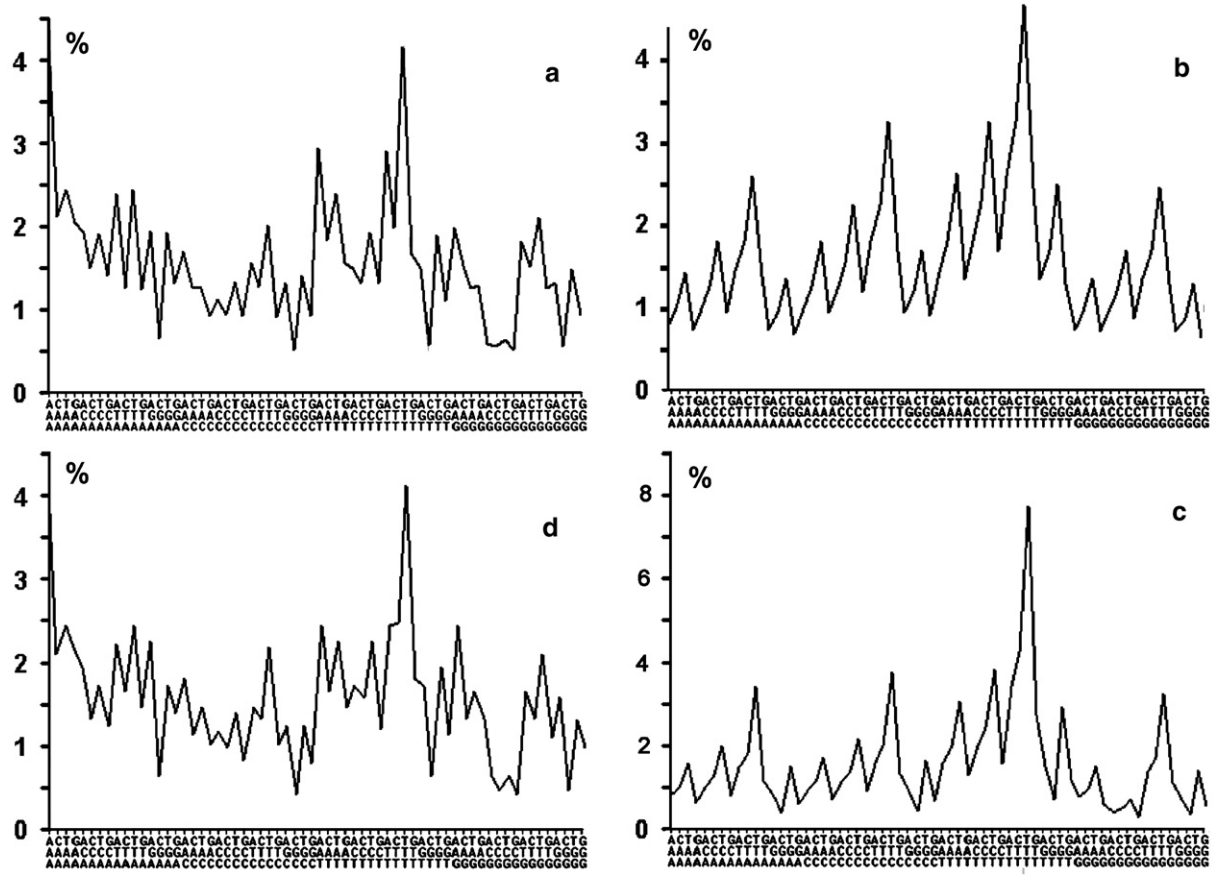

Fig. 3. Rendering of triplet profiles similar to the majority profile by the procedure described in the text (abscissa, triplets to be read from bottom to top; ordinate, fraction of triplets of entire genome). (a) The majority triplet profile [1,2]. (b) The stochastic expectation of the triplet profile resulting from an "urn experiment with replacement" (Eq. (2)) using the "initial" frequencies $p_0(A)=0.20$, $p_0(T)=0.36$, $p_0(C)=0.25$, $p_0(G)=0.19$. (c) Effect of replacing randomly 60% of all CG pairs with TT pairs on the "genome" with the triplet profile of (b). (d) Effect of 30,000 inversions/transpositions of 1-kb size on the triplet profile of the simulated genome of (c). This triplet profile is very similar to the majority profile of (a) (correlation coefficient is 0.957).

actually rather small. Using the base composition of Eq. (3), the stochastic-expectation genomes contained approximately 5% CG pairs on either strand. Therefore, the required sequence modification involved only $0.6\times5=3\%$ modification of their total dinucleotides. As a result, the triplet profiles that resulted from this modification (Fig. 3c) were qualitatively not very different from the initial profiles, although certain peaks were increased (note the change of scale in Fig. 3c). At this stage, the modified sequences still violated Chargaff's second parity rules.

I found no precedence for a CG-to-TT conversion in the literature. However, it may be interpreted as the result of the following two-step process, for which there is precedence in contemporary genomes. The first step may involve a CG-to-TA conversion, for which there is ample precedent in the literature, because methylated C is frequently converted into T [10]. Given a CG pair on (say) the Watson strand, base pairing requires that there is a CG pair opposite to it on the Crick strand (note that both strands are read in the 5′ to 3′ direction). Converting both (presumably methylated) C's into T's would generate a pair of mismatched TG dinucleotides. However, during the next round of replication they could be corrected into TA pairs [11,12].

The second step could be viewed as an A-to-T substitution that would turn some of the newly generated TA pairs on (say) the Watson strand into TT pairs. Correspondingly, their reverse complements on the Crick strand would turn into AA pairs. Such A to T substitutions have been observed in a number of cases [13]. Still, the above is merely a plausibility argument, since cytosine methylation is not a ubiquitous process. Therefore, it is quite possible, and perhaps even likely, that the postulated CG-to-TT conversion occurred by some other, yet to be discovered mechanism.

The intuitively obvious effect of numerous inversions and inverted transpositions

Subjecting the modified genomes to $N_{inv}=30{,}000$ inversions/transpositions (transposon size $\sigma=1$ kb) transformed their triplet profiles into the profiles shown in Fig. 3d. This profile appeared very similar to the majority profile (correlation coefficient 0.96) and fulfilled Chargaff's second parity rule (correlation coefficient between Watson and Crick strand 0.9994; Fig. 4d).

The action of 30,000 inversions/transpositions may seem rather obscure because it happens inside the impenetrable

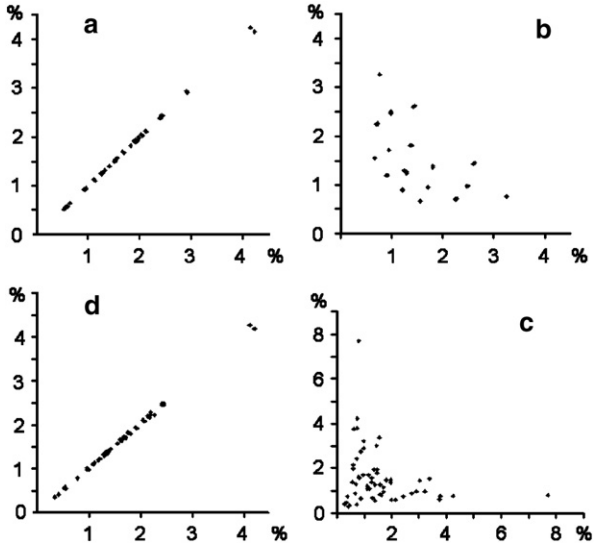

Fig. 4. Development of strand symmetry parallel to the generation of the majority-like triplet profile shown in Fig. 3. (a–d) Correlation plots between the triplet profiles of the Watson and Crick strands corresponding respectively to (a–d) of Fig. 3. (Abscissa, frequency of triplets on the Watson strand; ordinate, corresponding frequency of triplets on the Crick strand.) (a) Correlation plot of the majority triplet profile of Fig. 3a, which complies with strand symmetry quite accurately (correlation coefficient 0.9996). (b) Correlation plot of the stochastic-expectation triplet profile of Fig. 3b, which violates strand symmetry (correlation coefficient −0.5412). (c) Correlation plot of the triplet profile of Fig. 3c after the CG-to-TT conversion, which violates strand symmetry (correlation coefficient −0.1949). (d) Correlation plot of the final triplet profile of Fig. 3d after 30,000 inversions/transpositions that restore strand symmetry (correlation coefficient 0.9994).

"bowels" of a computer and, thereby, seems to offer little explanatory power. However, in this case the result is quite easy to understand even without the help of a computer.

As shown in the earlier article [1], the effect of a large number of inversions/transpositions is to equalize asymptotically the frequency of every triplet with the frequency of its reverse complement in an exponential fashion. For example, before the application of inversions/transpositions, the normalized frequency of the GAA triplet was $f_0(\mathrm{GAA})=0.6\%$ and that of its reverse complement was $f_0(\mathrm{TTC})=3.8\%$ (Fig. 5a). After a sufficiently large number of inversions/transpositions, both frequencies will became the same, namely $f_\infty(\mathrm{GAA})=f_\infty(\mathrm{TTC})=(0.6+3.8)/2=2.2\%$ (Fig. 5b). In other words, *the effect of the inversions/transpositions is nothing more than to generate the arithmetic means between the frequencies of triplets and their reverse complements*.

Denoting the reverse complement of a triplet △ by the symbol ▼, this result can be formulated as

$$f_\infty(\Delta)=f_\infty(\blacktriangledown)=[f_0(\Delta)+f_0(\blacktriangledown)]/2 \qquad (4)$$

The inversions/transpositions equalize not only the triplet frequencies, but the frequencies of mononucleotides, as well [1]. According to Eqs. (3) and (4), this would yield $p_\infty(\mathrm{A})=p_\infty(\mathrm{T})=[p_0(\mathrm{A})+p_0(\mathrm{T})]/2=28\%$ and $p_\infty(\mathrm{C})=p_\infty(\mathrm{G})=[p_0(\mathrm{C})+p_0(\mathrm{G})]/2=22\%$. Note that these frequencies differ from the frequencies $p_\mathrm{M}(X)$ of the majority profile. However, the difference is adjusted by the CG→TT conversions.

In view of this simple algorithm the accurate reproducibility of the effect of numerous inversions/transpositions is self-evident. Therefore, it seemed unnecessary to document the reproducibility of the results by statistical significance calculations. Nevertheless, I repeated the generation of the majority profile as described more than 20 times with identical results.

Parallel development of strand symmetry

Parallel to the generation of the majority-like triplet profile the described procedure also created almost perfect strand symmetry. Fig. shows the correlation plots between the triplet profiles of the Watson strand vs the Crick strand for each of the same stages as shown in Fig. 3. Since the majority profile complies with Chargaff's second parity rule, its correlation plot is a straight line (correlation coefficient 0.9996; Fig. 3a). The stochastic-expectation profile violates it before (correlation coefficient −0.5412; Fig. 3b) and after the application of a CG-to-TT conversion (correlation coefficient −0.1949; Fig. 3c). Consistent with our hypothesis [1], however, applying the large number of inversions/transpositions yielded not only a sequence whose profile was similar to the majority profile, but also one that fulfilled Chargaff's second parity rule (correlation coefficient 0.9994; Fig. 3d).

Note that the correlation plot is mirror symmetrical about the diagonal because, for each point $f(\Delta)$ on the abscissa plotted against $f(\blacktriangledown)$ on the ordinate, there is a symmetrically located point that plots $f(\blacktriangledown)$ on the abscissa against $f(\Delta)$ on the ordinate.

Parameter dependencies

The computed profiles did not match the majority profile exactly. However, their deviations from the majority profile were no larger than the discrepancies between the genomes of

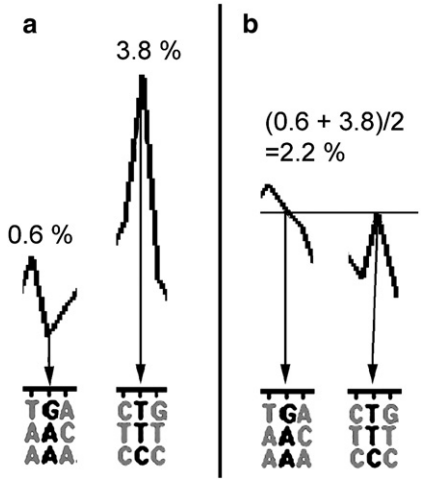

Fig. 5. Formation of the arithmetic mean between the initial frequency of each triplet and its reverse complement through numerous inversions/transpositions [1]. The example of the triplet AAG and its reverse complement CTT shown here was excised from Figs. 3c and 3d and printed to scale. (a) Initial frequencies of AAG and CTT before any inversions/transpositions. (b) Equalized final frequencies of the same two triplets after 30,000 inversions/transpositions representing the arithmetic mean of the initial values.

individual organisms within the majority group. Therefore, there was no reason to try to improve the matching any further by introducing more parameters. As a result, it appeared that the present derivation of the majority profile needed no more than two parameters, namely the initial values of p_0(A) and p_0(C). Either all other parameters were experimental data derived from the majority profile or the derivation did not need to vary them to match the majority profile.

To be sure, the above derivation contains six formal parameters. They include the initial AT content, the relationship $p_0(AT)=p_0(A)+p_0(T)$, and the initial values of p_0(A) and p_0(C). All other initial values of the base composition depended on these three. The remaining three parameters were the rate ρ of the conversion CG-to-TT, the number N_{inv} of inversions/transpositions, and the average size σ of the inverted transposons. Some of these parameters did not need to be varied or could be taken directly from the majority profile. For example, the number N_{inv} and size σ of inversions/transpositions had no further influence on the results once they exceeded the threshold values of 10,000 and 500, respectively (Figs. 6a and 6b). Likewise, the value of the conversion rate ρ had little influence on the results (Fig. 6c) in the range between 0.5 and 0.9. The range itself could be derived from a simple inspection of the stochastic-expectation genomes that had noticeably too few TT pairs and too many CG pairs relative to the majority distribution.

In view of this small number of relevant parameters, one may take the position that the described method of generating the majority profile is not a fitting procedure, but actually a way to calculate the initial base composition of the genomes of the majority class.

The other two classes of triplet profiles

In addition to the majority class, there were two other classes of triplet profiles, namely the minority class and the violator class [2]. The minority class consisted of GC-rich genomes with individually different profiles that complied with Chargaff's second parity rule, but were quite different from the majority profile. However, they could be turned into members of the majority class by a random conversion of a certain percentage of their G's and C's into equal numbers of A's and T's.

The violator class consisted of genomes whose profiles violated Chargaff's rules and were different from the majority profile. A subgroup of them, called moderate violators, differed only slightly from the majority class, as if relatively small mutations had changed them away from the majority profile.

The remaining violators, all mitochondrial genomes of the recent vertebrates, had surprisingly similar triplet profiles [2]. In the context of the present investigation it was found that their average profile could be described entirely by the stochastic expectation of Eq. (3) using their natural base frequencies of p_0(A)=33%, p_0(T)=26%, p_0(C)=28%, p_0(G)=13% (correlation with the stochastic-expectation profile=0.95). A subsequent exposure to the same inversions/transpositions and CG→TT conversion that were used in the above derivation of the majority profile was able to turn them into members of the majority class, as well. One may interpret these results as an indication that the two remaining classes represent genomes in transit toward or away from the majority class.

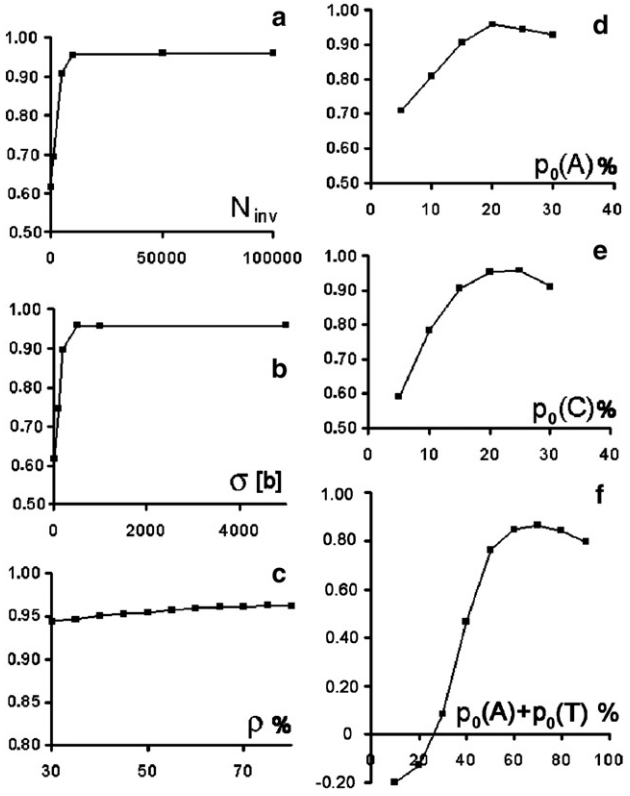

Fig. 6. Illustration of the typical effects of several basic parameters that influence the degree of matching between the triplet profiles of test genomes and the majority profile. (Abscissa, parameter values; ordinate, correlation coefficient between triplet profile and majority profile as a measure of the degree of matching.) The parameters were varied around the following "anchor" values: (1) the initial AT composition, $p_0(A)+p_0(T)=0.56$ (the corresponding value of the majority profile is 0.612); (2) the initial CG composition, $p_0(C)+p_0(G)=0.44$ (the corresponding value of the majority profile is 0.388); (3) the rate $\rho=0.6$ of the conversion CG-to-TT; (4) the number $N_{inv}=30{,}000$ of inversions/transpositions; and (5) the size $\sigma=1000$ [b] of the inverted transposons. In the depicted experiments four of the parameters were kept constant at the above values, while the fifth parameter was varied systematically. The results are shown in (a–d). (a) Degree of matching as a function of the number N_{inv} of inversions/transpositions. Once N_{inv} exceeded 10,000, the number of inversions/transpositions had no further influence on the results. (b) Degree of matching as a function of the size σ of the inverted transposons. Once σ exceeded 500, the size of the inversions/transposons had no further influence on the results. (c) Degree of matching as a function of the rate ρ of the conversion CG-to-TT. The conversion rate ρ yields a very shallow peak between 0.6 and 0.8, where it has only a minor effect on the result. (d) Degree of matching as a function of the initial base composition p_0(A). The values of $p_0(A)+p_0(T)=0.56$, $p_0(C)=0.25$, and $p_0(G)=0.19$ were kept constant while p_0(A) was varied. Peak value at $p_0(A)=0.20$. (e) Degree of matching as a function of the initial base composition p_0(C). The values of $p_0(A)=0.20$, $p_0(T)+p_0(G)=0.44$ were kept constant while p_0(C) was varied. Peak value at $p_0(C)=0.25$. (f) Degree of matching as a function of the initial AT content. The value of $p_0(G)=0.10$ was kept constant. Peak value at $p_0(A)+p_0(T)=0.70$.

Discussion

It may appear that the existence of a universal triplet profile imposes an unnecessary constraint on the evolution of genomes and organisms. However, this impression is misleading. This constraint on the triplet frequency has had as little effect on the variety of genomes as the well-known series of the most frequent three-letter words in the English language (the, and, for, are, but, not, you, all, any, etc.) has had on the English literature.

Such common features of a large body of texts, be they human texts or genomes, are no constraints on, but expressions of the results of the history of the underlying language. They may even offer advantages for communication. In this sense the present article suggests that the universal triplet profiles of genomes is the inevitable result of a common history of numerous inversions/transpositions that created this majority profile and that may offer advantages for horizontal gene transfer.

Of course, this does not exclude other mechanisms that may have contributed to the emergence of a universal triplet profile. For example, gene and genome duplication would certainly contribute to the universality of the profile. Obviously, the triplet profile of the concatenation of a duplicated genome is identical to its original profile (see the Appendix).

The inversion/transposition model of genome evolution

Both the present and other earlier results [1,2] are consistent with the notion that genome sequences arose randomly from a large pool of nucleotides, such as is postulated by the so-called "RNA world" hypothesis [5–9]. Whatever the mechanism of origin, the concept presented here assumes that initially a very large number of different nucleotide sequences existed, which became the raw materials for the future evolution of genomes. Yet, despite this large diversity of sequences, the base composition and the di-and trinucleotide (=triplet) profiles of these pregenomes were identical, as they reflected the nucleotide composition of the primordial soup from which they arose. Consistent with the ideas of the RNA world, the model also assumes that these initial pregenomes formed duplexes through extensive base pairing.

The present results lead to the novel idea that at these stages of evolution certain mechanisms of inversion, transpositions, and inverted transpositions began to operate extensively on the pregenomes and have acted on the subsequently evolving genomes ever since. As a result, genomes have developed the particular strand symmetry that was referred to as Chargaff's second parity rule [1]. At the same time these mechanisms also generated a common triplet profile for a majority of the evolving organisms.

Once generated, both of these genome properties were also maintained and stabilized by these mechanisms. This follows from the fact that the majority profile obeys Chargaff's second parity rule, which means that the Watson and Crick strands of majority genomes have the same triplet profiles. As a result, the inversion of a sufficiently large segment of such a genome carries on average the same set of triplets away from the Watson strand as it brings in from the Crick strand. In other words, once a genome had acquired the majority profile, further inversions/transpositions could change only its sequence, not its profile or the strand symmetry.

Although evolution proceeded to increase genome sizes more and more, it was not necessary to increase the frequency of inversions/transpositions any further. After all, most new genomes arose from precursors that already fulfilled Chargaff's second parity rule and that already had profiles similar to the majority profile. Only the newly added or modified sequences needed to be brought into compliance with the preexisting format of the genomes. Based on the results reported here, a few ten thousand inversions per several million years seem quite adequate to accomplish this task.

All the while a large percentage—although not a large number—of CG pairs converted to TT pairs. Although the above model had introduced this conversion as a next step after the stochastic-expectation genomes had formed, the results would have been the same had such conversions been continuously interspersed with the ongoing inversions/transpositions while accumulating to levels of up to 60%.

Do these many inversions/transpositions disrupt the genes?

From the point of view of gene evolution this model may paint a disturbing image. If hundreds of thousands of inversions/transpositions have recklessly altered the sequences of genomes, should they not have destroyed the evolving genes and the viability of the evolving organisms along with them?

Admittedly, inversions/transpositions probably did and still do destroy genes and create nonviable organisms on occasion. Selection, of course, was likely to eliminate them. However, since inversions/transpositions were relatively rare events, the survival of the species was never jeopardized because many contemporary organisms with unaltered genes continued to exist and reproduce. On the other hand, if inversions/transpositions happened to create novel genes and gene controls that offered a selective advantage for an organism [16], its offspring were likely to have outcompeted the others and driven evolution forward.

My data are consistent with the rarity of inversions/transpositions. The early genomes were presumably much shorter than the 8-Mb large test genomes used in my simulations and evolved into larger genomes in the course of tens to hundreds of millions of years later. As shown in the present study, only a relatively small number of some 10,000 inversions/transpositions were required to render the strands of their genome duplexes symmetrical and to create the majority profile. Thus the unification of this genome format required no more than an average of only 1 inversion/transposition per 1000 to 10,000 years.

Furthermore, genes seem to have evolved special ways to "protect" themselves. The underlying mechanisms are not understood, but genes seem to express a number of documented strategies to prevent the insertion of transposable elements into essential sequences, including the avoidance of promoters, preferential insertion into genes that exist in multiple copies, preferential insertion into introns that offer "safe" targets, preferential insertion into hot spots, and others [14,15].

Conservation of the majority profile by common mechanisms of genome variation

Another fundamental question about the feasibility of the above model of genome evolution concerns the effects of well-documented mechanisms of genome variation such as concatenations, recombinations, deletions, insertions, and point mutations. Would they not destroy the majority profiles of the evolving genomes?

As to point mutations, they involve only very small fractions of entire genomes and therefore do not alter the triplet profiles of the genome in any substantial way. As to the other mechanisms of genome variation, one can show that they actually preserve the triplet profiles of the genomes involved (see the Appendix).

Does the majority profile support horizontal gene transfer?

The evolutionary advantage of the existence of an almost universal genome format is not yet known. One may speculate, though, that it may facilitate horizontal gene transfer between vastly different species as the common format may have rendered native and foreign genes similar enough to be exchanged (see the Appendix). In this way it may speed up evolution considerably, as horizontal gene transfer makes it unnecessary for different organisms to rediscover the same beneficial genes many times over.

Materials and methods

The investigative computer program, dnaorg.exe, was written by G.A.-B. using Visual C+ (Microsoft, Redmond, WA, USA) and will be provided upon request.

Acknowledgments

I thank my wife Veena Prahlad (Northwestern University) and my friends and colleagues James Bartles and Alvin Telser (Feinberg School of Medicine, Northwestern University) for their criticism and patience during our many discussions about the presented subject.

Appendix A. The conservation of triplet profiles by the main methods of genome variation

Nomenclature

In the following, triplet profiles will be written symbolically as $f(\triangle)$, where the variable $\triangle$ represents all 64 triplets. If two segments of a genome sequence have the same triplet profile (e.g., the majority profile) they will be called "profile-related."

The "concatenation rule"

The concatenation of two profile-related segments preserves their triplet profile.

Assume two segments S_1 and S_2. Their concatenation leads to a sequence $S=S_1 \oplus S_2$. The counts $n(\triangle)$ for each triplet $\triangle$ of the Watson and Crick strands of the two duplexes and their concatenation will be denoted as $n_{\text{Watson}}(\triangle)_1$, $n_{\text{Crick}}(\triangle)_1$, $n_{\text{Watson}}(\triangle)_2$, $n_{\text{Crick}}(\triangle)_2$, and $n_{\text{Watson}}(\triangle)_{1\oplus2}$ and $n_{\text{Watson}}(\triangle)_{1\oplus2}$.

Assume S_1 and S_2 are profile-related:

$$f_{Watson}(\Delta)_1 = f_{Watson}(\Delta)_2 = f(\Delta) \tag{A1}$$

S_1 and S_2 will contain a total of n_1 and n_2 triplets, respectively. Hence their concatenation contains $m=n_1+n_2$ total triplets, to which the segment S_1 contributed $m_1(\triangle)=n_1 \cdot f(\triangle)$ and S_2 contributed $m_2(\triangle)=n_2 \cdot f(\triangle)$ triplets of the kind $\triangle$. Therefore, the concatenation contains

$$n_{1\oplus2}(\Delta) = m_1(\Delta) + m_2(\Delta) = (n_1 + n_2) \cdot f(\Delta) \tag{A2}$$

triplets of the kind $\triangle$.

On the other hand, by definition, the concatenation contains

$$\begin{aligned} n_{1\oplus2}(\Delta) &= m \cdot f_{Watson}(\Delta)_{1\oplus2} \\ &= (n_1 + n_2) \cdot f_{Watson}(\Delta)_{1\oplus2} \end{aligned} \tag{A3}$$

triplets of the kind $\triangle$.

After normalization Eqs. (A2) and (A3) yield the concatenation rule:

$$\begin{aligned} f_{Watson}(\Delta)_{1\oplus2} &= n_{1\oplus2}(\Delta)/m = f(\Delta) \\ &= f_{Watson}(\Delta)_1 = f_{Watson}(\Delta)_2 \end{aligned} \tag{A4}$$

Q.E.D.

Please note that at the very interface of concatenation of the two duplexes the profile symmetry and profile class membership may be violated for a stretch of one to two bases. Considering that both duplexes are assumed to be larger than 100 kb, this violation can be neglected. Nevertheless, the claim that the concatenation of two profile-related segments preserves their triplet profile should be qualified, in that it does not alter the profile in any substantial way. This more cautious formulation applies to all the following conclusions, but will not be repeated every time.

Invariance of triplet profiles against the actions of inversions/transpositions

In the following, all mentioned genome segments are assumed to have the same triplet profile, e.g., the majority profile.

(a) Deletions

Removing a segment D from a Watson strand generates a profile-related segment, S_1, in front and another, S_2, behind the deletion. After D is deleted, the two adjacent segments are concatenated and yield the duplex $S_d=S_1 \oplus S_2$. Both are profile-related as they belong to the same genome. According to the concatenation rule, their concatenation S_d preserves the triplet profile of the remainder.

(b) Insertions

If an insertion sequence I and a genome sequence S are profile-related, the insertion generates a concatenation of $S_i = S_1 \oplus I \oplus S_2$ with S_1 and S_2 being the segments of S before and after the insertion point. According to the concatenation rule the concatenated sequence S_i has the same triplet profile as its components.

(c) Transpositions

A transposition is the deletion of a segment T from a genome between two segments S_1 and S_2, followed by its reinsertion between two other segments S_1' and S_2' somewhere else. Since all segments involved are profile-related, sections (a) and (b) apply: Transpositions preserve the triplet profile of a genome.

(d) Inversions/inverted transpositions

An inversion is the transposition of an inverted segment T_i into its former location. If the genome in question complies with Chargaff's second parity rule, the Watson and Crick strands are profile-related. Therefore, an inverted segment T_i is a profile-related segment, and the result of section (c) applies: Inversions and inverted transpositions preserve the triplet profile of a genome.

(e) Horizontal gene transfer

If the genomes of two phylogenetically unrelated species, such as certain members of the majority class, have the same triplet profile, the same arguments that were used in sections (c) and (d) for transpositions within the same genome apply to horizontal gene transfer between those two species. In other words, horizontal gene transfer between the members of the majority class is a preserving operation, as well, even between phylogenetically unrelated species.

References

[1] G. Albrecht-Buehler, Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 17828–17833.

[2] G. Albrecht-Buehler, The three classes of triplet profiles of natural genomes, Genomics 89 (2007) 596–601.

[3] P.F. Baisnée, S. Hampson, P. Baldi, Why are complementary strands symmetric? Bioinformatics 188 (2002) 1021–1033.

[4] D. Mitchell, R. Bridge, A test of Chargaff's second rule, Biochem. Biophys. Res. Commun. 340 (2006) 90–94.

[5] The nature of modern RNA suggests a prebiotic RNA world, in: R.F. Gesteland, J.F. Atkins (Eds.), The RNA World, Cold Spring Harbor Monograph Series, vol. 24, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1993.

[6] W. Gilbert, The RNA world, Nature 31 (1986) 9618.

[7] T.R. Cech, A model for the RNA-catalyzed replication of RNA, Proc. Natl. Acad. Sci. U. S. A. 83 (1986) 4360–4363.

[8] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, S. Altman, The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme, Cell 353 (1983) 849–857.

[9] C. Woese, The Genetic Code, Harper and Row, New York, 1967.

[10] T. Lindahl, Instability and decay of the primary structure of DNA, Nature (London) 362 (1993) 709–715.

[11] D.A. Petrov, D.L. Hartl, Patterns of nucleotide substitution in *Drosophila* and mammalian genomes, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 1475–1479.

[12] E.B. Cambareri, B.C. Jensen, E. Schabtach, E.U. Selker, Repeat-induced G-C to A-T mutations in Neurospora, Science 244 (1989) 1571–1575.

[13] E.A. Drobetsky, A.J. Grosovsky, B.W. Glickman, The specificity of UV-induced mutations at an endogenous locus in mammalian cells, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 9103–9107.

[14] N.L. Craig, Target site selection in transposition, Annu. Rev. Biochem. 66 (1997) 437–474.

[15] G.-C. Liao, E.J. Rehm, G.M. Rubin, Insertion site preferences of the P transposable element in Drosophila melanogaster, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 3347–3351.

[16] B. McClintock, The significance of responses of the genome to challenge, Science 226 (1984) 792–801.