Genetica (2021) 149:81–88
https://doi.org/10.1007/s10709-021-00119-5

REVIEW

# Neutralism versus selectionism: Chargaff's second parity rule, revisited

**Donald R. Forsdyke**[1]

Received: 4 March 2021 / Accepted: 9 April 2021 / Published online: 20 April 2021


**Abstract**
Of Chargaff's four "rules" on DNA base frequencies, the functional interpretation of his second parity rule (PR2) is the most contentious. Thermophile base compositions (GC%) were taken by Galtier and Lobry (1997) as favoring Sueoka's neutral PR2 hypothesis over Forsdyke's selective PR2 hypothesis, namely that mutations improving local within-species recombination efficiency had generated a genome-wide potential for the strands of duplex DNA to separate and initiate recombination through the "kissing" of the tips of stem-loops. However, following Chargaff's GC rule, base composition mainly reflects a species-specific, genome-wide, evolutionary pressure. GC% could not have consistently followed the dictates of temperature, since it plays fundamental roles in both sustaining species integrity and, through primarily neutral genome-wide mutation, fostering speciation. Evidence for a local within-species recombination-initiating role of base order was obtained with a novel technology that masked the contribution of base composition to nucleic acid folding energy. Forsdyke's results were consistent with his PR2 hypothesis, appeared to resolve some root problems in biology and provided a theoretical underpinning for alignment-free taxonomic analyses using relative oligonucleotide frequencies (k-mer analysis). Moreover, consistent with Chargaff's cluster rule, discovery of the thermoadaptive role of the "purine-loading" of open reading frames made less tenable the Galtier-Lobry anti-selectionist arguments.

**Keywords** Base composition · Purine-loading · Speciation · Stem-loops · Taxonomy · Thermoadaptation

## Introduction

For molecular evolutionists, base composition (GC%), nucleic acid higher order structure and optimal temperature are topics of abiding interest. An exploration of their interrelationships by Galtier and Lobry (1997) had, by 2021, received an impressive number of citations (259) that were generally positive. They had argued that Chargaff's second parity rule (PR2) would be best interpreted in neutral terms as the result of "mutational bias" as had been proposed by Sueoka (1995). However, noting "two distinct interpretations," Galtier and Lobry (1997) wrote:

> The second interpretation is that PR2 is the result of a "selection pressure favoring mutations that generate complementary oligonucleotides in close proximity, thus creating a potential to form stem-loops" (Forsdyke 1995a). According to this hypothesis, deviations from PR2 are deviations from the ideal case, where the whole genome is involved in secondary structures.

Simply stated, Galtier and Lobry favored a neutral explanation for PR2 over Forsdyke's selectionist explanation. Rather than proving Sueoka's proposal, their paper, through "analyzing the effect of temperature on the G + C content of bacterial genomes" set out to disprove Forsdyke:

> We have looked for this effect in molecules known to be involved in secondary structures (tRNAs, 5S rRNAs, and the stems of 16S and 23S rRNAs) and in the whole genome. If the second selectionist hypothesis is correct, a high proportion of the genomes should be involved in forming secondary structure, so genomic G+C content should follow the same pattern as that of the folded RNA G+C content.

As long argued by Bernardi (1993) and later recognized by Lobry and Sueoka (2002), the premise that the base compositions (GC%) of thermophile genomes, and of the corresponding mRNAs which function primarily as *templates*, would need to adapt to high temperatures in the same manner as tRNAs and rRNAs, which function primarily

✉ Donald R. Forsdyke
forsdyke@queensu.ca

[1] Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L3N6, Canada

by virtue of their *structures*, was incorrect. Prompted by a recent commentary (Meyer 2021), after clarification of terminology relating to Chargaff's PR2, and description of structure analysis methodology that relates to Chargaff's GC rule (Forsdyke and Mortimer 2000), the selectionist case is here updated.

## Chargaff's second parity rule terminology

Sueoka (1995) referred to the equimolar pairing of purines and pyrimidines in duplex DNA (Watson and Crick 1953) as in accordance with the "*inter*strand base-pairing rule (BPR)" (my italics). However, he then went on to distinguish *two* forms of *intra*strand parity, that he termed "PR1" and "PR2":

> "This article presents two types of intrastrand parity rules: the type 1 parity rule (PR1) is concerned with base substitution rates within one strand of DNA, and the type 2 parity rule (PR2) is concerned with the base composition at equilibrium within one strand of DNA."

Subsequently these terms came to be applied differently. Sueoka's BPR has become known as "Chargaff's first parity rule" (PR1). The base equivalence in single-stranded DNA that Chargaff later discovered, has become known as "Chargaff's second parity rule" (PR2). The numbering reflects discovery chronology, not the number of strands interacting. Given the assumption that PR2 equivalences resemble PR1 as indicating equimolar pairing (A with T and G with C), it follows that PR2 has implications for the possible structures that a single-stranded nucleic acid, be it RNA or DNA, can adopt (Forsdyke and Mortimer 2000).

## Nucleic acid structure, speciation and Chargaff's GC rule

Taxonomists construct phylogenetic trees that model species diversification. Organisms showing similar characters are placed close together. Organisms showing different characters are placed at greater distances. Some characters are more useful for tree construction than others. When nucleic acid sequences became available, they were used for tree construction. The closer two sequences, the closer were considered the corresponding organisms. To count the number of base differences between two sequences, various methods for comparative alignments of long strings of bases were introduced. However, the approach was empirical and did not consider the possibility that some aspects of sequences, other than base strings, might better relate to the underlying evolutionary processes that had perhaps initiated, or otherwise fostered, divergence. For example, an early indication of language divergence is a difference in accents. In this case, lining up long texts would necessitate the inclusion of much redundant information. Some measure of accent difference should better display the relationship between the emerging languages because redundancies would be eliminated.

Chargaff recognized that the quantity of G + C relative to that of A + T (expressed as GC%) was a species characteristic—his "GC rule" (Forsdyke and Mortimer 2000). Just as accent or dialect affects a spoken text in its entirety, so base *composition* (GC%) tends to uniformity either genome-wide or in large genomic sectors (Bernardi 1993). Many differences in base composition that accumulate between species over time are selectively neutral. These *interspecies* mutations influence genome-wide oligonucleotide (k-mer) frequencies, of which base composition (GC%) is an indicator (see later). Changes in these frequencies could, in turn, influence the "kissing" interactions between the loops of stem-loop structures (Kleckner and Weiner 1993), so acting to generate, and/or sustain, members of emerging species by preventing recombination with parental forms (Forsdyke 1996, 2014, 2019a, b). Thus, measurement of base composition (GC%) is a simple method of eliminating the influence of base order that, for displaying a fundamental relationship between species, is redundant information.

However, sometimes the reverse is required. Just as a local arrangement of words conveys specific meaning to a text, so base *order* should better reflect local non-uniformity (variation) *within* members of a species. To clarify this, a method for eliminating the influence of base composition is needed. Since the energetics of the folding of a single-stranded nucleic acid into a stem-loop structure depend on *both* the composition and order of its bases, a localized sequence (e. g. a 200 base "window" in the sequence) that is rich in the strongly-pairing bases G and C, will tend to have a stable structure simply by virtue of its base composition, rather than of its unique base order. This high GC% value can obscure the contribution of the base order-dependent component of the folding energy, which provides a sensitive indicator of *local intraspecies* pressures for the conservation of function within a population (i.e. a mutated organism is eliminated by natural selection so no longer can be assayed for function in the population). Thus, elimination of the base composition-dependent component should facilitate focus on this local folding.

In studies of RNA virus structure, Le and Maizel (1989) compared calculated structures of natural RNA sequences with the structures of the same sequences that had been shuffled to randomize base order (i.e. the base order-dependent component of the folding energy in the natural sequence was eliminated). They found that base-randomized sequences generally had weaker folding energy values and, on these grounds, they concluded that the natural folding was

statistically (and perhaps biologically) significant. However, by subtracting values for shuffled sequences from those for the natural sequence, the Le-Maizel methodology permitted a distinction between the contributions of base composition (a genome-wide function) and base order (a local function). Thus, with a pipeline between the various programs that were offered by the Wisconsin Genetics Computer Group, the base composition and base order-dependent components were separated and individually assessed ("folding of randomized sequence difference" analysis; FORS-D analysis; Forsdyke 1995a, b, 2013, 2014; Xu et al. 2007; Zhang et al. 2008a).

In practice (Fig. 1), a window of 200 bases is moved stepwise along a natural sequence. A folding program (Zuker 1989) is applied to the sequence in each window to obtain "folding of natural sequence" (FONS) values for each window, to which both base composition and base order will have contributed. The four bases in each sequence window are then shuffled to destroy their order while retaining base composition. The folding energy is then again determined. This shuffle-and-fold "Monte Carlo" procedure is repeated ten times and the average (mean) folding value is taken as the "folding of randomized sequence mean" (FORS-M) value for that window. This reflects the contribution of base composition alone. The base order-dependent component is

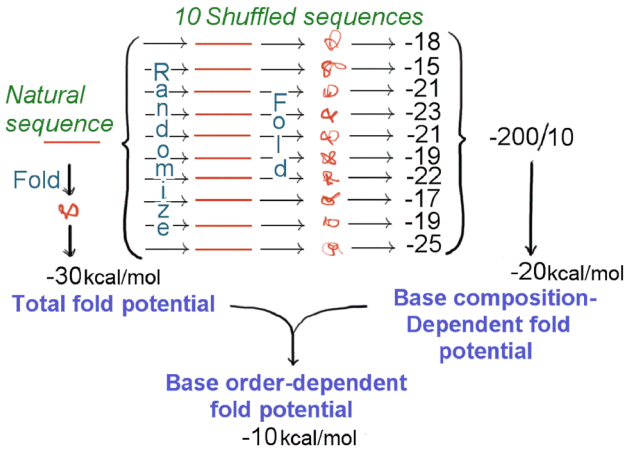


**Fig. 1** Determination of the base order-dependent component of stem-loop (fold) potential by subtracting the base composition-dependent component from total stem-loop potential. A natural sequence (horizontal red line at left) when optimally folded (vertical arrow at left) is calculated to have a certain stability (e. g. − 30 kcal/mol). Its base order is then randomized to produce ten shuffled sequences that share only their base compositions with the originating natural sequence. These are then optimally folded to obtain corresponding stability values. Idiosyncrasies, due to the base order that each randomized sequence has acquired due to the shuffling, are averaged out (at right) to determine the contribution of base composition to the total fold potential. The contribution of base order is determined by subtraction. This figure is with permission reproduced from Forsdyke (2016)

then derived by subtraction from the FONS value. This is the "folding of randomized sequence difference"(FORS-D) value. Local fluctuations in FONS profiles of genomes are mostly due to changes in the FORS-D component, whereas the FORS-M component (base composition), while making a major contribution to folding energetics, is relatively constant (Zhang et al. 2008b).

This approach was employed by others who, rather than shuffling the *four* bases, favored retaining some base order information (Workman and Krogh 1999). Accordingly, they shuffled groups of bases (e. g. the *sixteen* dinucleotides). Following disparagement of the conceptual basis of four base shuffling, which was duly clarified (Forsdyke 2007a), the validity of single base level shuffling is now generally accepted and is being applied routinely to viral genomes (Witteveldt et al. 2014; Andrews et al. 2018; Simmonds 2020). The Monte Carlo procedure can also be simplified to decrease FORS-M computational time (Chen et al. 1990; Washietl et al. 2005). Software ("Bodslp") written by Jiansheng Wu (Zhang et al. 2008a), retains the Monte Carlo approach and was developed by Shungao Xu as "Random Fold-Scan" for Windows-based systems (Xu et al. 2007).

In addition to assisting the study of infectious viruses and protozoa (Xue and Forsdyke 2003), FORS-D analysis proved fruitful when applied to topics such as speciation (Forsdyke 1996; 2014; Zhang et al. 2008b), the origin of introns (Forsdyke 1995b, c, 2013), relating structure to recombination breakpoints and deletions (Zhang et al. 2005a, b), use of a single sequence (rather than alignments) for the determination of positive Darwinian selection (Forsdyke 2007b), and showing that a highly conserved region in HIV-1 genomes associates with an RNA packaging signal (Forsdyke 1995d). The latter is now seen as a potential "Achilles heel" (Ingemarsdotter et al. 2018), so encouraging a similar approach to SARS-CoV-2 (Zhang and Forsdyke 2020).

## Update on selectionism: roles of higher K-mers

At the time of the paper of Galtier and Lobry (1997), the neutralist-selectionist controversy was beginning to swing in favor of selectionism, at least among geneticists (Hey 1999):

> The findings on codon bias, the rediscovery and relevance of the Hill–Robertson effect and the fact that many loci reveal nonneutral local patterns of variation suggest that natural selection is highly pervasive at the DNA level. To a scientist bred on neutrality, the discoveries of recent years present some daunting theoretical and empirical challenges. A rejection of neutrality is not the same thing as an understanding of natural selection, and the discoveries mean that

evolutionary genetics has become more difficult than it once seemed.

Indeed, the handling editor of the Galtier-Lobry paper, Giorgio Bernardi, supported the selectionist viewpoint (D'Onofio et al. 1999; Bernardi 2000):

> The low GC levels of some thermophilic bacteria do not contradict, as claimed (Galtier and Lobry 1997), the selectionist interpretation ... . Indeed, different strategies were apparently developed by different organisms to cope with long-term high body temperatures. It is now known that the DNAs of such thermophilic bacteria are very strongly stabilized by particular DNA-binding proteins (Robinson et al. 1998) and that, in turn, their proteins can be stabilized by thermostable chaperoninins (Taguchi et al. 1991).

A further discussion of base composition in thermophilic organisms, noting especially their purine-enrichment—in keeping with Chargaff's cluster rule—was provided by Forsdyke and Mortimer (2000), Hurst and Merchant (2001) and Lambros et al. (2003). These affirmed that a positive correlation between (G + C)% and optimum growth temperature would mainly apply to RNAs whose primary function was structural (ribosomal and transfer RNAs) and would *not* apply to mRNAs and the corresponding genomic DNA sequences from which they had been transcribed. Lobry and Sueoka (2002) conceded the possibility of error:

> On the other hand, one may assume that some [base] positions are not free to deviate from PR2 because of selective pressure for some function. For instance, there are 100 copies per enterobacterial genome of palindromic sequences in intergenic spaces [Bachellier et al. 1999], which may be under selective pressure to preserve their palindromic character and therefore follow PR2 (as pure palindromic sequences are effectively base-paired).

In later papers Lobry moved further from neutral theory (Necşulea and Lobry 2006) and also recognized a possible selective basis for purine-enrichment: "Compared to mesophilic species, thermophilic genomes are significantly enriched in purines and in purine-clusters" (i.e. local violations of PR2), for which "a possible explanation" is "the existence of a selective pressure to avoid undesirable RNA–RNA interactions" (Lobry and Necşulea 2006), as had been suggested by Cristillo et al. (1998). Thus, finding that: "The behavior of the most discriminating codon AGG with respect to G + C content is especially puzzling," they reported that "relatively high AGG frequencies prevent us from excluding the hypothesis of a selective pressure in favor of this codon …. Since AGG is a pure-purinic codon, the latter hypothesis may be linked to the observed purine enrichment in thermophilic and hyperthermophilic species." In this regard it was suggested that those seeking to associate certain amino acids with the high stability of thermophile proteins should consider the possibility that those corresponding to purine-rich codons might be mere placeholders (Forsdyke 2015a; Bize et al. 2021).

Despite Lobry's concession, various papers (Forsdyke 2002, 2015b; Forsdyke and Bell 2004), and a textbook treatment (Forsdyke 2016), there is no general agreement on the functional implications of PR2. As part of the 50th anniversary celebration of the Journal of Molecular Evolution, the topic was addressed by Meyer (2021), who noted that "Increasingly it appears that G + C content in genomes may be the result of a combination of neutral and selection processes that are quite subtle (Reichenberger et al. 2015)." Meyer's commentary aimed to "present the motivation behind Galtier and Lobry's original paper and the larger questions addressed by the work, how these questions have evolved over the last two decades, and the impact of Galtier and Lobry's manuscript in fields beyond these questions." Thus, their paper was not chosen for comment simply because it refuted the selectionist viewpoint. It was cited as a helpful resource concerning the influence of optimum growth temperature (OGT) on the GC% of certain RNA species:

> Despite the specific nature of the hypothesis addressed, the two [major] findings for which this paper is most frequently cited are quite general. The first is the lack of relationship between OGT and genomic G + C content. The second is that G + C content in the stems of the 16S and 23S rRNAs, and generally in the 5S rRNA and tRNAs, does correlate with organismal OGT."

Indeed, Galtier and Lobry (1997) cited Forsdyke (1995a, b) which, together with a paper the following year (Forsdyke 1996), provide a grounding for several of "the larger questions" as will be summarized here. Thus, Meyer (2021) mentioned three applications of their two major findings:

> The two major findings of Galtier and Lobry have spurred significant further work that encompasses a range of different applications that take advantage of the relationships between OGT, structured RNA G + C content, and genomic G + C content. These include: prediction of organism OGT based on 16S rRNA sequence, separation or enrichment of DNA extracted from microbial communities for a particular sub-populations based on G + C content, and computational methods for structured RNA identification."

Two of these three applications relate to the work cited by Galtier and Lobry. A grounding for computational methods for structured RNA identification was, as has been discussed above, provided by Forsdyke (1995b). The separation

of DNA from community subpopulations, currently known as "metagenomics" (alignment-free k-mer analysis; Forsdyke 2019a; Bize et al. 2021), is grounded on analysis of relative oligonucleotide frequencies (Forsdyke 1995a). The latter paper began by noting that, for example, a high GC% species would have more GC-rich oligonucleotides (e. g. GCG, GAG, GCC) than AT-rich oligonucleotides (e. g. ATA, AGA, ATT). Intriguingly, even if they had *equivalent* GC% values, the k-mer patterns of phylogenetically close taxa were more similar than those of distant taxa. Would a specific oligonucleotide pattern be a consequence of a primary evolutionary pressure for a high GC%, or vice versa? Which was the cart and which was the horse—mononucleotide or oligonucleotide? Essentially the same question was posed by Blake et al. (1992) when considering the context-dependence of point mutations with special reference to immediate upstream and downstream bases (i.e. k-mers where k = 2 or 3):

> The classic studies of Benzer (1961), showing the sites of … genes of T4 mutate with different efficiencies, could be seen as reflecting different neigbor effects at the several steps in the fixation of mutations. Benzer designated sites of high efficiency as hotspots. ... An influence of the neighbor environment could also be partially responsible for the biases in neighbor frequences in sequences (Josse et al. 1961). Whether such bias is a cause or an effect of one or more of these events remains to be demonstrated … .

A primary role for higher k-mers was argued on theoretical grounds (Forsdyke and Bell 2004), which were supported by statistical analyses of human genomes that indicated a k-mer optimum of at least 7, and were held to reveal principles that, although undetermined, were "fundamental" (Aggarwala and Voight 2016):

> Although the underlying mechanisms that determine how nucleotide sequences change over time remain to be addressed, we posit that the features identified from our model provide important clues in elucidating these fundamental principles."

Along similar lines, Morozov (2017) reported a k-mer optimality, ranging from k = 5 to k = 7, among a wide variety of species and pondered:

> Thus, there is no question of whether *k*-mer distribution … is species-specific or whether the divergence of these distributions correlates with evolutionary distances. However, there is no answer to *why* it does.

Within-species recombination and between-species anti-recombination would seem to be the selective principle they seek (Forsdyke 1996, 2014, 2019a, b; Bozdag et al. 2021). Supporting recombination, natural selection seeks out higher order k-mers, and the quantity of 1-mers (GC%) is secondary to this. Meyer (2021) rightly asked for the "causes for G + C content variability: neutral processes or natural selection?" A biological grounding for this was provided in terms of primarily *neutral* variation that can generate k-mer differences that inhibit recombination and initiate sympatric speciation (reproductive isolation) prior to the involvement of natural selection in the speciation process. Yet, following Chargaff's cluster rule, the *selective* influence of "R-loading" (base skew in favor of A and G; Cristillo et al. 1998) has the potential to locally (in open reading frames) impact GC% to the extent that a reciprocal relationship between GC% and AG% can emerge (Lao and Forsdyke 2000; Mortimer and Forsdyke 2003).

Indeed, citing Bell and Forsdyke (1999); Meyer (2021) noted that "sequences that are actively transcribed also tend to display purine loading." This creates a base-skew in exons which violates PR2 more than in introns. Hence there is usually more DNA secondary structure potential in introns, especially in genes under positive Darwinian selection (Forsdyke 1995c, 2007b). Consistent with this, in a study of genome k-mer frequencies, Bultrini et al. (2003) showed that the second parity rule is followed more closely in intronic and intergenic DNA than in exonic DNA. They concluded:

> A very interesting feature of the *C. elegans* intron vocabulary is its being almost entirely composed of pairs of reverse complementary oligos. ... A symmetrical trend is apparent on a scale of a few kilobases in individual *C. elegans* introns. This short-range property of introns is not simply due to their symmetrical base composition, since it is drastically reduced in randomized introns. Rather, it results from the preferred use of reverse complementary oligomers ... . It would be tempting to link the above symmetry properties of introns to formation of stem-loop structures.

As for mechanism, Meyer (2021) suggested: "The most satisfying explanations for the maintenance of Chargaff's second rule invoke frequent duplication, inversion, and transposition events in the genome" (Albrecht-Beuhler 2006). These may indeed favor PR2 *maintenance* of a base relationship that could, however, have originated in the need for recombination-based error-correction that would be expected to have arisen very early in the evolution of living forms ("introns early;" Forsdyke 1995b, 2013).

## Concluding remarks

What Chargaff first described as "regularities" in base composition have since become referred to as his four "rules" – PR1, PR2, the GC rule and his "cluster" rule (Forsdyke and Mortimer 2000; Forsdyke 2016). Among these, functional

interpretation of PR2 has proved the most confusing and contentious (see Online Resource 1). Galtier and Lobry (1997) portrayed the problem in stark terms as a contest between neutralist and selectionist viewpoints, centering their case around the papers of Sueoka (1995) and Forsdyke (1995a). They did not clarify Sueoka's terminology and for several years many may have assumed that their thermophile data had settled the issue. Happily, a quarter century later Meyer (2021) reopened the issue for a journal anniversary, touching on the history of a contest between two groups: The *neutralists*—Galtier, Lobry and Sueoka, with Motoo Kimura (1924–1994) of "neutral theory" fame (Hey 1999), close by. The *selectionists*—Bernardi and Forsdyke and their various coworkers.

Simply stated, Sueoka, whose early studies offered profound insights (Forsdyke 2016, 2019a), was wrong on this issue; so perhaps we should now be more open to selectionist interpretations. The thermoadaptation issue has been laid to rest. Now the focus is on how "larger questions" that were "ultimately the subject of Galtier and Lobry's paper" have "evolved over the last two decades." Meyer pondered:

> It is still unclear whether G + C content variation may be generated by neutral processes such as mutational bias or biased gene conversion, or is primarily the result of natural selection. Furthermore, even if such variation is the result of natural selection, is selection acting on the genomic DNA itself, or rather on the molecules (e.g. RNAs and proteins) encoded by the DNA?

The answer appears to be that unlike, for example, improving the utilization of a substrate such as glucose that cannot adapt to improve its metabolism by specific enzymes, DNA *can* adapt to assist the function of enzymes acting upon it. Thus, *both* substrate and enzymes acting on that substrate can be targeted by natural selection to improve recombination efficiency—hence genome-wide stem-loop potential. Recombination begins with higher order k-mers in loops seeking complementary k-mers on other loops. Failure to find that complementation can initiate speciation.

Meyer considered the strength of the Galtier-Lobry paper was that it "ultimately seeded several other fruitful areas of research," and spurred interest in "far reaching" questions. The primary conclusion of Galtier and Lobry (1997) was that:

> Our results do not support the notion that selection pressure induces complementary oligonucleotides in close proximity and therefore numerous secondary structures in prokaryotic DNA, as the genomic G+C content does not behave in the same way as that of folded RNA with respect to optimal growth temperature."

Although this may be questionable, some of their findings do support a selectionist PR2 viewpoint. Concerning the loops of stem-loop structures and optimum growth temperature (Topt) they noted: "a weak correlation between the G + C content of 16S and 23S rRNA loops and Topt …. It may be due to … a tertiary structure effect." This implies loop-loop interactions that would require equimolar pairing, *at a distance*, of purines and pyrimidines, so *adding* to the *local* compliance of stems with PR2. Furthermore, Galtier and Lobry (1997) agreed that "there are also thermophilic species with low genomic G + C contents, such as *Pyrococcus furiosus* (Topt: 97 °C, G + C%: 38; Fiala and Stetter 1986), that indicate no selective advantage of a high genomic G + C content at high temperature." Indeed, Bernardi (1993) had considered "the thermal stabilization of genomes might be due not to an increase in G + C but to other physiological adaptations,".

A hopefully growing consensus viewpoint can be described in the following general terms. Whereas RNA molecules transcribed from DNA may primarily function *either* structurally (e. g. rRNA, tRNA) *or* as templates for protein synthesis (mRNA), this division of labor is absent from their genomic source. *Entire* DNA molecules have both structural and templating functions that must sometimes compete for sequence space—a feature that could explain some properties of introns (Forsdyke 1995b, c, 2013).

The function of a nucleic acid depends both on which bases it contains (*composition*) and on their *order*. A distinction can be made since the contribution of base composition to DNA structural information (stem-loop potential) tends to be relatively uniform and genome-wide and can be equated energetically with the mean of several shuffled versions. On the other hand, specific templating information tends to be irregular and localized, and can best be studied energetically without the contribution of base composition. The contribution of base order to structural functions can be assessed as the difference between the calculated structure of a natural nucleic acid with that of the corresponding randomized (bases shuffled) version (Fig. 1).

The *genes* for rRNAs have close sequence similarities with those of the rRNAs they encode, which have been shaped by natural selection acting primarily in the cytoplasm. Thus, for the genome locations of these rRNA genes, the DNA structure is strongly under cytoplasmic influence. But rRNA genes are few and far between. They are not representative of entire genomic sequences that encode a multiplicity of dispersed, widely varying, RNAs—including mRNAs—the structures of which are less likely to be under cytoplasmic influence. Thus, mRNA structures primarily reflect the potential of the corresponding DNA sequences to adapt to favor recombination (Kleckner and Weiner 1993).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10709-021-00119-5.

**Acknowledgements** I thank Michelle Meyer for her choice of anniversary topic and advice. YouTube hosts two videos on Chargaff's rules (https://www.youtube.com/watch?v=QS6dqEwI9R8; https://www.youtube.com/watch?v=8iCgSrCwQko). Further background is hosted by the Internet Archives "Wayback Machine": https://wayback.archive-it.org/7641/20200423143429/http:/post.queensu.ca/~forsdyke/bioinfor.htm#Neutralism%20Versus%20Selectionism. A preprint is posted in the Genomes and Genetics section of *SSRC-BioRN Computational Biology*.

## Declarations

**Conflict of interest** The author declares that there is no conflict of interest.

## References

Aggarwala V, Voight BF (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet 48:349–355

Albrecht-Beuhler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci USA 103:17828–17833

Andrews RJ, Roche J, Moss WN (2018) ScanFold: an approach for genome-wide discovery of local RNA structural elements – applications to Zika virus and HIV. PeerJ 6:e6136

Bachellier S, Clement JM, Hofnung M (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. Res Microbiol 150:627–639

Bell SJ, Forsdyke DR (1999) Deviations from Chargaff's second parity rule correlate with direction of transcription. J Theor Biol 197:63–76

Benzer S (1961) On the topography of the genetic fine structure. Proc Natl Acad Sci USA 47:403–415

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17

Bize A, Midoux C, Mariadassou M, Schbath S, Forterre P, Da Cunha V (2021) Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. BMC Genomics 22:186

Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol 34:189–200

Bozdag GO, Ono J, Denton JA, Karakoc E, Hunter N, Leu J-Y, Greig D (2021) Breaking a species barrier by enabling hybrid recombination. Curr Biol 31(R16):1-R185

Bultrini E, Pizzi E, Del Guidice P, Frontali C (2003) Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. Gene 304:183–192

Chen J-H, Le S-Y, Shapiro B, Currey KM, Maizel JV (1990) A computational procedure for assessing the significance of RNA secondary structure. Comput Appl Biosci 6:7–18

Cristillo AD, Lillicrap TP, Forsdyke DR (1998) Purine-loading of EBNA-1 mRNA avoids sense-antisense "collisions." FASEBJ 12:A1453

D'Onofrio G, Jabbari K, Musto H, Alvarez-Valin F, Cruveiller S, Bernardi G (1999) Evolutionary genomics of vertebrates and its implications. Ann New York Acad Sci 870:81–94

Fiala G, Stetter KO (1986) *Pyrococcus furiosus* sp. nov. represents a new genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch Microbiol 145:56–61

Forsdyke DR (1995a) Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41:573–581

Forsdyke DR (1995b) A stem-loop "'kissing"' model for the initiation of recombination and the origin of introns. Mol Biol Evol 12:949–958

Forsdyke DR (1995c) Conservation of stem-loop potential in introns of snake venom phospholipase $A_2$ genes: an application of FORS-D analysis. Mol Biol Evol 12:1157–1165

Forsdyke DR (1995d) Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. J Mol Evol 41:1022–1037

Forsdyke DR (1996) Different biological species "broadcast" their DNAs at different (G+C)% "wavelengths." J Theoret Biol 178:405–417

Forsdyke DR (2002) Symmetry observations in long nucleotide sequences. a commentary on the discovery note of Qi and Cuticchia. Bioinformatics 18:215–217

Forsdyke DR (2007a) Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. J Theor Biol 248:745–753

Forsdyke DR (2007b) Positive Darwinian selection. Does the comparative method rule? J Biol Sys 15:95–108

Forsdyke DR (2013) Introns first. Biol Theor 7:196–203

Forsdyke DR (2014) Implications of HIV RNA structure for recombination, speciation, and the neutralism-selectionism controversy. Mic Infect 16:96–103

Forsdyke DR (2015a) Purine loading as a thermal adaptation. PubMed Commons 26254668 stored at Hypothes.is: https://hypothes.is/search?q=tag%3APubMedCommonsArchive+forsdyke

Forsdyke DR (2015b) Neutral theory not supported. PubMed Commons 9169555 stored at Hypothes.is: https://hypothes.is/search?q=tag%3APubMedCommonsArchive+forsdyke

Forsdyke DR (2016) Evolutionary Bioinformatics, 3rd edn. Springer, New York

Forsdyke DR (2019) Hybrid sterility can only be primary when acting as a reproductive barrier for sympatric speciation. Biol J Linn Soc 128:779–788

Forsdyke DR (2019) Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation. Biol J Linn Soc 128:239–250

Forsdyke DR, Bell SJ (2004) Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl Bioinformatics 3:3–8

Forsdyke DR, Mortimer JR (2000) Chargaff's legacy. Gene 261:127–137

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44:632–636

Hey J (1999) The neutralist, the fly and the selectionist. Trends Ecol Evol 14:35–38

Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 268:493–497

Ingemarsdotter CK, Zeng J, Long Z, Lever AML, Kenyon JC (2018) An RNA-binding compound that stabilizes the HIV-1 gRNA packaging signal structure and specifically blocks HIV-1 RNA encapsidation. Retrovirol 15:25

Josse J, Kaiser AD, Kornberg A (1961) Enzymic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem 236:864–871

Kleckner N, Weiner BM (1993) Potential advantages of unstable interactions for pairing of chromosomes in meiotic, somatic and premeiotic cells. Cold Spring Harb Symp Quant Biol 58:553–565

Lambros RJ, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles 7:443–450

Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res 10:228–236

Le S-Y, Maizel JV (1989) A method for assessing the statistical significance of RNA folding. J Theor Biol 138:495–510

Lobry JR, Necşulea A (2006) Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. Gene 385:128–136

Lobry JR, Sueoka N (2002) Asymmetric directional mutational pressures in bacteria. Genome Biol 3(research0058):1

Meyer MM (2021) Revisiting the relationships between genomic G + C content, RNA secondary structures, and optimal growth temperature. J Mol Evol 89:165–171

Morozov AA (2017) *k*-mer distributions of aminoacid sequences are optimised across the proteome. bioRxiv. https://doi.org/10.1101/190280

Mortimer JR, Forsdyke DR (2003) Comparison of responses by bacteriophage and bacteria to pressures on the base composition of open reading frames. App Bioinf 2:47–62

Necşulea A, Lobry JR (2006) Revisiting the directional mutation pressure theory: The analysis of a particular genomic structure in *Leishmania major*. Gene 385:28–40

Reichenberger ER, Rosen G, Hershberg U, Hershberg R (2015) Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol 7:1380–1389

Robinson H, Gao Y-G, Bradford SM, Edmondson SP, Shriver JW, Wang AH-J (1998) The hyperthermophile chromosomal protein Sac7d sharply kinks. Nature 392:202–205

Simmonds P (2020) Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses. MBio 11:e01661-e1720

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Taguchi H, Konishi J, Ishii N, Yoshida M (1991) A chaperonin from a thermophylic bacterium, *Thermus thermophylus*, that controls refoldings of several thermophilic enzymes. J Biol Chem 266:22411–22418

Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102:2454–2459

Watson JD, Crick FH (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171:737–738

Witteveldt J, Blundell R, Maarleveld JJ, McFadden N, Evans DJ, Simmonds P (2014) The influence of viral RNA secondary structure on interactions with innate host cell defences. Nucleic Acids Res 42:3314–3329

Workman C, Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res 27:4816–4822

Xu SG, Wei JF, Zhang CY (2007) A FORS-D analysis software "Random_fold_scan" and the influence of different shuffle approaches on FORS-D analysis. J Jiangsu Univ (Med Ed) 17(461–466):470

Xue HY, Forsdyke DR (2003) Low complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. Mol Biochem Parasitol 128:21–32

Zhang C, Forsdyke DR (2020) Potential Achilles heels of SARS-CoV-2 displayed by the base order-dependent component of RNA folding energy. https://europepmc.org/article/PPR/PPR229852

Zhang C-Y, Wei J-F, He S-H (2005) The key role for local base order in the generation of multiple forms of China HIV1 B'/C intersubtype recombinants. BMC Evol Biol 5:53

Zhang C-Y, Wei J-F, He S-H (2005) Local base order influences the origin of *ccr5* deletions mediated by DNA slip replication. Biochem Genet 43:229–237

Zhang C, Xu S, Wei J-F, Forsdyke DR (2008) Microsatellites that violate Chargaff's second parity rule have base order-dependent asymmetries in the folding energies of complementary DNA strands and may not drive speciation. J Theor Biol 254:168–177

Zhang C-Y, Wei J-F, Wu J-S, Xu W-R, Sun X, He S-H (2008) Evaluation of FORS-D analysis: a comparison with the statistically significant stem-loop potential. Biochem Genet 46:29–40

Zuker M (1989) Computer prediction of RNA secondary structure. Meth Enzym 180:262–289