Validity of the 2nd Chargaff rule and Inversion Symmetry.

David Horn

Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel

Email: horn@tau.ac.il.

Abstract

There exist interesting identities of nucleotide counts on single chromosomal strands. The first of this kind was the 2nd Chargaff rule (SCR), later generalized to Inversion Symmetry (IS). These rules can nowadays be investigated using generative statistical tools, leading to novel observations. They imply the existence of empirical IS for nucleotide strings of length k, where k grows logarithmically with the length of the chromosomal strand. On the other hand they can be used to point out slight deviations from exact SCR which are correlated with different numbers of gene counts on the two chromosomal strands.

Erwin Chargaff has made, in 1950, the important observation that the numbers of nucleotides in DNA satisfy #A=#T and #G=#C [1]. This statement, made on the basis of biochemical observations with fairly large errors, played a crucial role in realizing that DNA has a two strand structure with base-pair binding, as proposed by Crick and Watson in their seminal paper [2].

The second Chargaff rule (SCR) [3] states that the same sets of identities among counts of nucleotides hold for each long enough single DNA strand. But whereas the 1st rule can in hindsight be justified by base-pair binding, the SCR has become a curious puzzle and stayed relatively unknown. Over 50 years have passed since the publication of SCR. It has since been verified by numerous investigators on chromosomes of many species [4] and found to fail only for mitochondria, plasmids, single-stranded DNA viruses and

RNA viruses. Moreover, it has been generalized to an Inversion Symmetry (IS) rule, stating that any string S of k nucleotides (e.g. S=ACTG, with k=4) occurs on a single strand approximately the same number of times as its inverse (reverse and transpose) string S^{inv} (e.g. S^{inv}=CAGT, while the reverse would be S^{rev}=GTCA). This has been first observed [4] for low values of k, and is now known to hold up to k=10 on long human chromosomes [5,6,7], with relative errors $\epsilon(S)=|N(S)-N(S^{inv})|/(N(S)+N(S^{inv}))$, which are on average (over all S of length k) less than 10%.

In a recent paper [7] we have analyzed the questions whether, in addition to measuring empirically the accuracy of IS in data, one can also formulate it as a rule within a probabilistic generative model (answer: yes)? and can one then validate or refute such rules (surprises lie ahead)? if significant deviations from the rule occur, can they be correlated with other biological phenomena (yes)? and how do these rules come into being (evolution of course…)?

To move from empirical observations to probabilistic rules, we viewed [7] the number of empirical observations, N, as instances of a Poisson variable ***N,*** befitting stochastic occurrences of the string S on a long chromosome which are independent of one another. Inversion symmetry is then formulated as $\boldsymbol{N}(S)=\boldsymbol{N}(S^{inv})$, meaning that S and S^{inv} have identical probability distributions. This can be tested by asking whether the empirical measurements N(S) and $N(S^{inv})$ on a chromosomal strand of length L agree with the expectations of the IS Poisson model. The latter can be judged by testing the distribution (over all S of length k) of

$$Z(S)=(N(S)-N(S^{inv}))/(\,N(S)+N(S^{inv}))^{½}$$

which should be standard normal (Gaussian with variance=1). The results of this study indicated the significance of our findings and are displayed in Fig. 1 (for IS of k=8 strings) and in Fig. 2 (showing that reverse symmetry is not valid).

Our analysis revealed a dichotomy between “accuracy of empirical IS” and “validity of the IS Poisson model” or “significance of IS breaking”: For k=1

to 4, accuracy is high (small error) but the strict rule is invalid. This means that one observes small discrepancies of IS and SCR, with an average $\epsilon(S)$ of order 10^{-3}, which are statistically significant and may indicate the existence of biological reasons for departure from the symmetry. For large k, the accuracy of empirical IS diminishes, but the validity of the rule cannot be refuted. Both are due to the increasing number of different strings S of the same length k, which grows like 4^k, and the corresponding lower values of counts N(S) for each one of these strings. It also turns out that, if one fixes the allowed error of empirical IS at a given margin (e.g. $\epsilon<10\%$) then the largest k for which empirical IS holds, grows proportionally to log(L). The latter turns out to be a valid universal description of empirical data of many species and of chromosomal sections of varying lengths [7] (see figure 3).

It is known [8] that, within genes, there often exists a compositional asymmetry on the coding strand with an excess of #T+#G>#A+#C, which may be relevant for the operation of the transcription machinery. We observed [7] that the breaking of SCR on large chromosomes correlates well with an existing asymmetry of gene counts on the two strands of the chromosome; moreover, nucleotide count asymmetries agree for most chromosomes with the gene compositional asymmetry. This demonstration can be carried out, so far, for the human genome only, where sufficient data allow for drawing these conclusions.

Finally we are left with the question how SCR and IS came into being. A reasonable answer [4,6,9] is that this is due to the development of chromosomes throughout evolution, which is known to involve reordering of chromosomal sections, leading to syntheny maps [10] between chromosomes of different species. Since rearrangements are implemented in both directions of the chromosome, large numbers of random rearrangements lead to the observed phenomena.

In summary, both SCR and its generalization into Inversion Symmetry (IS), are valid biological rules. On SCR one notices a small violation of the rule, which correlates well with a small asymmetry of gene occurrences on the two strands. These rules may be viewed as emerging phenomena, caused by

the tinkering of chromosomal evolution with chromosomal sections, rearranging them randomly in either a direct or inverted fashion into a novel DNA molecule.

References

1. Chargaff E . Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 1950, 6(6):201-9.
2. Crick F and Watson JD . Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 1953, 171: 737-738.
3. Rudner R, Karkas JD, Chargaff E . Separation of B. subtilis DNA into complementary strands. III. Direct Analysis. *Proc Natl Acad Sci USA* 1968, 60:921-922.
4. Mitchell D, Bridge R. A test of Chargaff's second rule. *Biochem Biophys Res Commun,*2006, 340(1):90-94; Albrecht-Buehler G. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverse transpositions. *Proc Natl Acad Sci USA* 2006, 103 (47) 17828-17833; Prabhu, V V. Symmetry observations in long nucleotide sequences. *Nuc. Acids Res.* 1993, 21 (12): 2797-2800
5. Baisnee, P-F, Hampson, S and Baldi, P. Why are reversary DNA strands symmetric? *Bioinformatics* 2002, 18, 1021-1033.
6. Kong S-G, Fan W-L, Chen H-D, Hsu Z-T, Zhou N, Zheng B, and Lee H-C. Inverse symmetry in complete genomes and whole-genome inverse duplication. *PlosOne* 2009, 4, e7553.
7. Shporer, S, Chor, B, Rosset, S, Horn, D. Inversion symmetry of DNA k-mer counts: validity and deviations. BMC Genomics 2016, 17:696
8. Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program; Green ED. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Gen.*2003, 33: 514-517.
9. Okamura K, Wei J, Scherer S W. Evolutionary implications of inversions that have caused intra-strand parity in DNA. *BMC Genomics*. 2007; 8: 160
10. Pevzner P, Tesler G. Genome rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes. *Genome Res*. 2003,13: 37-45.

Figures selected from our publication [7]:

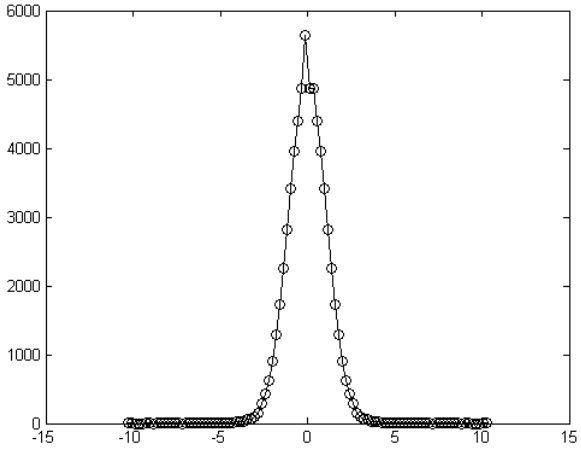

Fig. 1. Z-distribution of all strings of k=8 on human chromosome 1 agree well with the expected standard normal distribution.

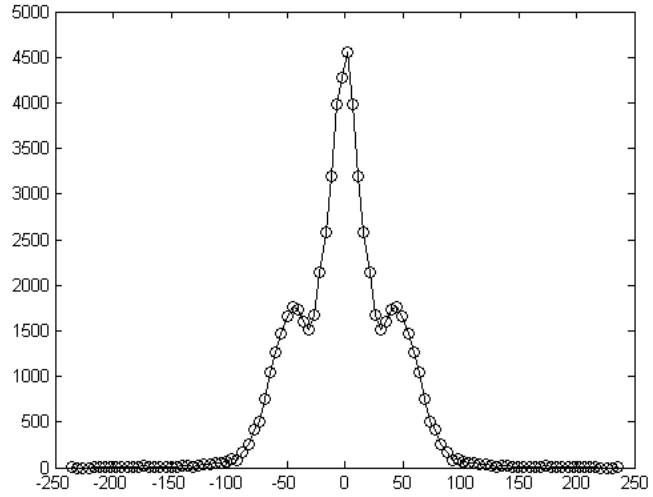

Fig. 2. Z-distribution of reverse pairs of k=8 strings, i.e. where S^{inv} is replaced by S^{rev}, shows a completely different behavior on human chromosome 1. This indicates that, while inverse symmetry is valid, reverse symmetry is not.

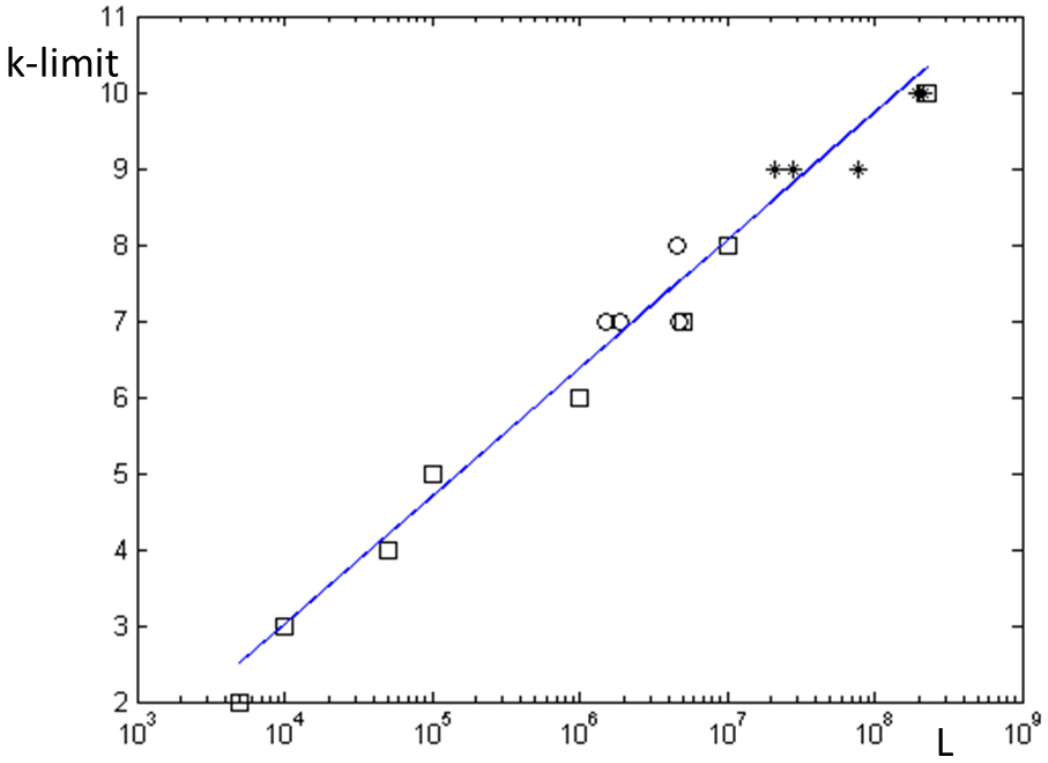

Fig. 3: k-limits (largest values of k for which the average $\epsilon(S)$ <10%), are plotted *vs* chromosomal length L. The figure displays a universal logarithmic behavior growing like 0.73 ln(L), which agrees with the expectation of the IS Poisson model. Boxes are human data (both full chromosomes and chromosome sections of various lengths), stars denote examples of other eukaryotes, and circles represent examples of prokaryotes.