Atlas of Science
another view on science
http://atlasofscience.org

Is inversion symmetry of chromosomes a law of nature?

Chargaff has made, in 1950, the important observation that the numbers of nucleotides in DNA satisfy #A=#T and #G=#C. This played a crucial role in realizing that DNA has a two strand structure with base-pair bindings (of A to T and C to G) as proposed by Crick and Watson.

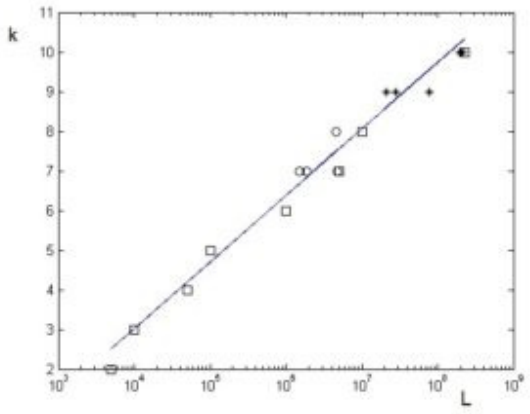

Fig. 1. Largest k-values, for which the IS empirical error is <10%, are plotted vs chromosomal length L. The line grows like 0.73 ln(L), which is the expected result of the IS Poisson model. Boxes are human data (full chromosomes as well as sections of various lengths), stars denote examples of other eukaryotes, and circles represent examples of prokaryotes. This figure is reproduced from Shporer et al.

Another paper, in 1968, has revealed the second Chargaff rule (SCR) stating that the same sets of identities 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 remained a curious puzzle. It has been verified by numerous investigators on chromosomes of many species and 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). This is now known to hold up to k=10 on long human chromosomes with relative errors less than 10%.

To move from empirical observations, such as number N of counts of some string on a chromosome, to probabilistic rules, one should view N as an instance of a Poisson variable ***N***, befitting stochastic occurrences of the string S which are independent of one another. Inversion symmetry is then formulated as $\boldsymbol{N}(S)=\boldsymbol{N}(S^{inv})$. 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.

Our analysis revealed a dichotomy between "accuracy of empirical IS" and "significance of IS breaking": For k=1 to 4, accuracy is high but the strict rule is invalid, i.e. there exist statistically significant small discrepancies. For large k, the accuracy of empirical IS diminishes, but the validity of the rule cannot be refuted. It also turns out that, if one fixes the allowed error of empirical IS at a given margin (e.g. 10%) then the largest k for which it holds, grows proportionally to log(L). The latter turns out to be a valid universal description of empirical data (Fig. 1).

It is known 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. The breaking of SCR on large chromosomes may be related to this compositional asymmetry: it turns out to correlate well with another small asymmetry, that of gene counts on the two strands of human chromosomes; moreover, nucleotide count asymmetries agree for most chromosomes with the gene compositional asymmetry.

Finally we are left with the question how SCR and IS came into being. A reasonable speculation is that this is due to the development of chromosomes throughout evolution, which is known to involve reordering of chromosomal sections. 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 small violations, which correlate with a small asymmetry of gene occurrences on the two strands. The IS rules may be viewed as emerging phenomena, caused by the tinkering of evolution with chromosomal sections, rearranging them randomly in either a direct or inverted fashion into novel DNA molecules.

David Horn
Sackler School of Physics and Astronomy, Tel Aviv University, Israel

Publication

Inversion symmetry of DNA k-mer counts: validity and deviations.
Shporer S, Chor B, Rosset S, Horn D
BMC Genomics. 2016 Aug 31