Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2025, 13(12), 4490-4497
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

The Entropy of a DNA Strand

Pirooz Mohazzabi[1*], Nathan Hugh Jones[1], Riley Houston Tucker[1], Nicholas James Winter[2]

[1]Department of Mathematics and Physics, University of Wisconsin-Parkside, Kenosha, WI, USA
[2]Department of Biological Sciences, University of Wisconsin-Parkside, Kenosha, WI, USA
Email: *mohazzab@uwp.edu

How to cite this paper: Mohazzabi, P., Jones, N.H., Tucker, R.H. and Winter, N.J. (2025) The Entropy of a DNA Strand. *Journal of Applied Mathematics and Physics*, **13**, 4490-4497.
https://doi.org/10.4236/jamp.2025.1312246

Received: November 17, 2025
Accepted: December 21, 2025
Published: December 24, 2025

Abstract

Based on the number of available sites for adenine, guanine, thymine, and cytosine on all DNA strands, the information theoretical entropy as well as the configurational entropy of a DNA strand is calculated. The number of possibilities for the formation of human DNA is also discussed and it is shown that, in addition to sharing 99.9% of their DNA, there must be other hidden factors limiting the observed genetic variations in humans.

Keywords

Entropy, DNA, Nitrogenous Bases, Configuration, Information

1. Introduction

A DNA molecule is a double helix consisting of two strands of chains of atoms, each made of sugar (deoxyribose) and phosphate groups, which alternate to form each strand. Four nitrogenous bases, adenine (A), cytosine (C), guanine (G), or thymine (T) are attached to each sugar molecule, forming a nucleotide, which contains the genetic information of the species. The ratio of the number of four bases in a DNA strand can vary. However, it is expected that each base has the same probability to attach to the strand.

Entropy, which is a measure of disorder, randomness, or uncertainty, is a fundamental concept in many areas of science, including physics, chemistry, and biology. There are three definitions of entropy; thermal entropy, configurational entropy, and information theoretic entropy.

Thermal entropy, which is a fundamental concept in thermodynamics and statistical mechanics, is defined for a reversible process by [1]-[3]

$$dS_{th} = \frac{dQ}{T} \quad \text{(reversible process)} \tag{1}$$

where S is the entropy, Q is the heat absorbed by the system, and T is the

absolute temperature. This entropy is a measure of thermal randomness of the particles of the system.

Configurational entropy, which is a measure of geometrical disorder of particles of the system, is defined by [4]

$$S_{con} = k_B \ln \Omega \tag{2}$$

where $k_B = 1.381 \times 10^{-23}$ J/K is the Boltzmann constant. Here Ω is the number of microstates consistent with macroscopic properties of the system.

Information theoretic entropy, associated with the number of possible outcome of a random variable, is defined by [5]

$$S_{inf} = -\sum_{i=1}^{n} p_i \ln p_i \tag{3}$$

where p_i is the probability of the outcome i. If the base of the logarithm is chosen to be 2, this entropy is referred to as Shannon entropy. The configurational entropy and information theoretic entropy are closely related. This is because for an event with equally probable outcomes we have

$$p_i = \frac{1}{\Omega} \tag{4}$$

for all values of i. Therefore,

$$S_{inf} = -\sum_{i=1}^{n} \frac{1}{\Omega} \ln \frac{1}{\Omega} = \sum_{i=1}^{n} \frac{1}{\Omega} \ln \Omega \tag{5}$$

But because $n = \Omega$ and the terms in the summation are all equal,

$$S_{inf} = \Omega \left(\frac{1}{\Omega} \ln \Omega \right) = \ln \Omega \tag{6}$$

Therefore, except for the Boltzmann constant, S_{inf} and S_{con} are the same.

The entropy of DNA has been studied by many investigators from various points of view. For example, Zhang *et al.* [6] developed a new algorithm to study segmenation of DNA sequences using quadratic divergence. Zhang [7] used Shannon entropy and genome order index to segment DNA sequences. Sherwin [8] discussed advantages of entropy-based genetic diversity measures at levels from gene expression to landscapes. Kirillova [9] calculated topological and metric entropies of the DNA sequences for different organisms. Schmitt and Herzel [10] presented a method to estimate higher order (or block) entropies for DNA sequences when the actual number of observations is small compared with the number of possible outcomes. Vinga and Almeida [11] used continuous approach to extend Shannon's formalism to DNA sequences. Loewenstern and Yianilos [12] indicated that under certain assumptions, DNA would have a much lower entropy than expected otherwise. However, they stated that surprisingly this has not been the case for many natural DNA sequences, including portions of the human genome. Koslicki [13] states that of all the entropy-theoretic notions, topological entropy has been the most difficult to implement due to various reasons. He then defines a new approximation to topological entropy to avoid the existing difficul-

ties. His approximation shows that the entropy of introns is significantly higher than that of exons, contrary to the previous calculations. Privalov and Crane-Robinson [14] investigated folding/unfolding DNA duplexes of various size and composition by superprecise calorimetry, and revised several long-standing beliefs regarding the forces responsible for the formation of the double helix. These are just a few articles regarding the entropy of a DNA strand. Many others can be found in the literature, including those related to engineering and drug applications [15] [16]. Nevertheless, none of these references have calculated the entropy of a DNA strand from a simple phenomenological point of view.

In what follows, we evaluate the entropy of a DNA strand from a different, yet simple and straightforward perspective, which is closely related to the number of ways that genetic codes can be embedded in the molecule.

2. The Entropy of a DNA Strand

We consider a DNA strand consisting of N sites to which the four bases can be attached. Let αN of these sites be occupied by adenine (A), βN of them occupied by cytosine (C), γN occupied by guanine (G), and the rest occupied by thymine (T). Here $0 \le \alpha, \beta, \gamma \le 1$, but such that $\alpha + \beta + \gamma \le 1$.

The number of ways that adenine can occupy the N sites is given by [17]

$$C_{\alpha N}^{N} = \frac{N!}{(N-\alpha N)!(\alpha N)!} = \frac{N!}{\left[(1-\alpha)N\right]!(\alpha N)!} \tag{7}$$

Then the number of the remaining sites available to cytosine is $(1-\alpha)N$. The number of ways that cytosine can occupy these sites is

$$C_{\beta N}^{(1-\alpha)N} = \frac{\left[(1-\alpha)N\right]!}{\left[(1-\alpha)N-\beta N\right]!(\beta N)!} = \frac{\left[(1-\alpha)N\right]!}{\left[(1-\alpha-\beta)N\right]!(\beta N)!} \tag{8}$$

Finally, the number of sites available to guanine is $(1-\alpha-\beta)N$, and the number of ways that guanine can occupy these sites is

$$C_{\gamma N}^{(1-\alpha-\beta)N} = \frac{\left[(1-\alpha-\beta)N\right]!}{\left[(1-\alpha-\beta)N-\gamma N\right]!(\gamma N)!} = \frac{\left[(1-\alpha-\beta)N\right]!}{\left[(1-\alpha-\beta-\gamma)N\right]!(\gamma N)!} \tag{9}$$

The rest of the available sites on the strand are occupied by thymine, which is only one way that it can be done.

According to the foregoing discussion, and since the configurations (7), (8), and (9) are independent, the total number of configurations that the four bases can be attached to the strand is given by the product rule [18],

$$\Omega = C_{\alpha N}^{N} C_{\beta N}^{(1-\alpha)N} C_{\gamma N}^{(1-\alpha-\beta)N} \tag{10}$$

Substituting for $C_{\alpha N}^{N}$, $C_{\beta N}^{(1-\alpha)N}$, and $C_{\gamma N}^{(1-\alpha-\beta)N}$ from Equations (7)-(9), we get

$$\Omega = \frac{N!}{(\alpha N)!(\beta N)!(\gamma N)!\left[(1-\alpha-\beta-\gamma)N\right]!} \tag{11}$$

Evaluating $\ln\Omega$, using Stirling approximation $\ln n! = n(\ln n - 1)$ for $n \gg 1$,

and simplifying, we obtain

$$\ln\Omega = -N\left[\alpha\ln\alpha + \beta\ln\beta + \gamma\ln\gamma + (1-\alpha-\beta-\gamma)\ln(1-\alpha-\beta-\gamma)\right] \quad (12)$$

which is the information theoretic entropy of the DNA strand. The configurational entropy is simply this quantity multiplied by the Boltzmann constant, as stated earlier.

To find the values of α, β, and γ that make the entropy a maximum or minimum, we set the partial derivatives of the entropy with respect to these parameters equal to zero,

$$\frac{\partial}{\partial\alpha}\ln\Omega = -N\ln\left(\frac{\alpha}{1-\alpha-\beta-\gamma}\right) = 0 \quad (13)$$

or

$$\frac{\alpha}{1-\alpha-\beta-\gamma} = 1 \quad (14)$$

Similarly, we get

$$\frac{\beta}{1-\alpha-\beta-\gamma} = 1 \text{ and } \frac{\gamma}{1-\alpha-\beta-\gamma} = 1 \quad (15)$$

From the three Equations (14) and (15), we find

$$\alpha = \beta = \gamma = \frac{1}{4} \quad (16)$$

It is straightforward to show that these values correspond to the maximum of the DNA entropy. Although it is not possible to plot a four-dimensional graph to show this, we can provide a simple graph as a visual aid. Let α be the fraction of one of the nucleotide bases, and consider the quantity $\ln\Omega/N$ as a function of α as α varies between 0 and 1, keeping each of the other fractions equal to $\frac{1-\alpha}{3}$. This reduces Equation (12) to

$$\frac{\ln\Omega}{N} = -\alpha\ln\alpha - (1-\alpha)\ln\left(\frac{1-\alpha}{3}\right) \quad (17)$$

A graph of this function is shown in Figure 1. According to this graph, clearly $\ln\Omega/N$ becomes a maximum when $\alpha = 1/4$.

3. Numerical Values of the Entropy

With the values of the fractions α, β, and γ, each equal to 1/4, Equation (12) gives a maximum entropy of

$$S_{inf} = \ln\Omega = (2\ln 2)N = 1.3863N \quad (18)$$

where N is the total number of sites available on each DNA strand for the four bases. This number varies greatly depending on the organism and the specific gene within the organism, but for the human genome (haploid), this number is approximately 3.2×10^9 for all strands (with twice this number for the double helix) [19]-[22]. Therefore, the maximum information theoretic entropy for

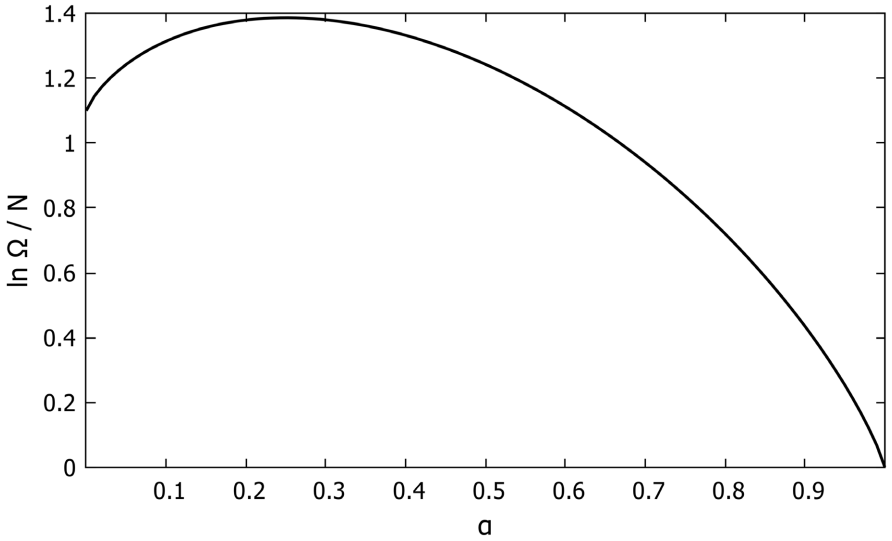

Figure 1. A graph of $\ln\Omega/N$ as a function of α, *i.e.*, Equation (17).

human DNA is approximately 4.436×10^9. The minimum entropy is clearly zero, which is the case when all the sites are occupied by only one of the bases, say adenine, with $\alpha=1$, $\beta=\gamma=0$.

It should be pointed out that the configurational entropy of the double helix is the same as that of a single strand. The reason is that in the formation of double helix, adenine and thymine bond together, and cytosine and guanine bond together. Therefore, the structure of the second strand in the double helix is dictated by the first strand, consequently no new configurational uncertainty is added to the system.

As stated earlier, when the probabilities of the outcomes of an event are all equal, the configurational entropy and the information theoretical entropy are the same except for the Boltzmann constant according to Equation (2). Therefore, the maximum configurational entropy of all DNA strands is

$$\begin{aligned} S_{con} &= k_B \ln\Omega = 1.381\times10^{-23}\left(4.436\times10^9\right) \\ &= 6.126\times10^{-14}\ \text{J/K} = 3.824\times10^5\ \text{eV/K} \end{aligned} \tag{19}$$

Therefore, the maximum configurational entropy of all DNA strands is about 0.382 MeV/K.

Going back to the information theoretical entropy, since its maximum for a human DNA is 4.436×10^9, the minimum probability that a given strand would have a specific configuration is

$$p_{\min} = \frac{1}{\Omega_{\max}} = \frac{1}{\mathrm{e}^{4.436\times10^9}} = \frac{1}{10^{1.927\times10^9}} \tag{20}$$

Therefore, the minimum probability of a given DNA strand having a specific configuration is unimaginably small. The maximum probability is obviously equal to unity, corresponding to $\ln\Omega=0$.

4. Discussion

When there are equal numbers of adenine, cytosine, guanine, and thymine in a genome, *i.e.* when each of the nucleotide bases occupy 1/4 of the N available sites, the information theoretic entropy and hence the configurational entropy of the DNA becomes a maximum of $\ln\Omega_{\max} = 1.3863N$. Then the maximum number of possibilities for the genetic code would be equal to $\Omega_{\max} = e^{1.3863N}$. Depending on the value of N, this result is applicable to all living species, including animals, plants, bacteria, and viruses. For humans, with $N = 3.2\times10^9$, after changing the base of the exponential we obtain $\Omega_{\max} = 10^{1.927\times10^9}$, which is an unimaginably large number of possibilities for DNA.

It is generally believed that all humans share about 99.9% of their DNA, allowing only 0.1% for the genetic variations [23] [24]. Some other authors have suggested 99.63%, allowing 0.37% for genetic variations [25]. Using the former values, since 0.1% of the available sites on a human DNA strand is 3.2×10^6, the resulting number of possibilities for genetic variations is $\Omega_{\max} = 10^{1.927\times10^6}$, which is still unimaginably large. Consequently, assuming that the total number of available sites, $N = 3.2\times10^9$, reported in the literature is correct, there must be additional constraints that drastically reduce the number of available sites for human genetic variations, such as mutation, selection, and genetic drift over evolutionary history.

As **Figure 1** shows, depending on the values of α, β, and γ, the factor of N in Equation (12) can vary from a maximum of 1.386 to zero. Therefore, the total number of genetic possibilities depends on these ratios and can be much lower than the values mentioned above. Based on Chargaff's rule [26], it is estimated that the composition of the nucleotide bases in a human DNA should be about 30% for adenine and thymine, and 20% cytosine and guanine. Nevertheless, even with these values, the factor of N in Equation (12) is 1.37, which is still too high. This is an issue that should be looked into. However, it is not the intention of this research to investigate this discrepancy, but rather to bring it to the attention of researchers in the field.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Callen, H.B. (1985) Thermodynamics and an Introduction to Thermostatistics. 2nd Edition, Wiley, 36.

[2] Borgnakke, C. and Sonntag, R.E. (2013) Fundamentals of Thermodynamics. 8th Edition, Wiley, 263.

[3] Huang, K. (1987) Statistical Mechanics. 2nd Edition, Wiley, 14.

[4] Wikipedia, Configuration Entropy. https://en.wikipedia.org/wiki/Configuration_entropy

[5] Wikipedia, Entropy (Information Theory).

https://en.wikipedia.org/wiki/Entropy_(information_theory)

[6] Zhang, C., Gao, F. and Zhang, R. (2005) Segmentation Algorithm for DNA Sequences. *Physical Review E*, **72**, Article ID: 041917. https://doi.org/10.1103/physreve.72.041917

[7] Zhang, Y. (2009) Relations between Shannon Entropy and Genome Order Index in Segmenting DNA Sequences. *Physical Review E*, **79**, Article ID: 041918. https://doi.org/10.1103/physreve.79.041918

[8] Sherwin, W.B. (2010) Entropy and Information Approaches to Genetic Diversity and Its Expression: Genomic Geography. *Entropy*, **12**, 1765-1798. https://doi.org/10.3390/e12071765

[9] Kirillova, O.V. (2000) Entropy Concepts and DNA Investigations. *Physics Letters A*, **274**, 247-253. https://doi.org/10.1016/s0375-9601(00)00557-0

[10] Schmitt, A.O. and Herzel, H. (1997) Estimating the Entropy of DNA Sequences. *Journal of Theoretical Biology*, **188**, 369-377. https://doi.org/10.1006/jtbi.1997.0493

[11] Vinga, S. and Almeida, J.S. (2004) Rényi Continuous Entropy of DNA Sequences. *Journal of Theoretical Biology*, **231**, 377-388. https://doi.org/10.1016/j.jtbi.2004.06.030

[12] Loewenstern, D. and Yianilos, P.N. (1999) Significantly Lower Entropy Estimates for Natural DNA Sequences. *Journal of Computational Biology*, **6**, 125-142. https://doi.org/10.1089/cmb.1999.6.125

[13] Koslicki, D. (2011) Topological Entropy of DNA Sequences. *Bioinformatics*, **27**, 1061-1067. https://doi.org/10.1093/bioinformatics/btr077

[14] Privalov, P.L. and Crane-Robinson, C. (2018) Translational Entropy and DNA Duplex Stability. *Biophysical Journal*, **114**, 15-20. https://doi.org/10.1016/j.bpj.2017.11.003

[15] Zhang, D.Y., Turberfield, A.J., Yurke, B. and Winfree, E. (2007) Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA. *Science*, **318**, 1121-1125. https://doi.org/10.1126/science.1148532

[16] Breslauer, K.J., Remeta, D.P., Chou, W.Y., Ferrante, R., Curry, J., Zaunczkowski, D., *et al.* (1987) Enthalpy-Entropy Compensations in Drug-DNA Binding Studies. *Proceedings of the National Academy of Sciences*, **84**, 8922-8926. https://doi.org/10.1073/pnas.84.24.8922

[17] Rozanov, Y.A. (1969) Probability Theory: A Concise Course. Revised English Edition, Translated by R.A. Silverman, Dover, 7.

[18] Brualdi, R.A. (2010) Introductory Combinatorics. 5th Edition, Pearson Education, Inc., 47.

[19] International Human Genome Sequencing Consortium (2004) Finishing the Euchromatic Sequence of the Human Genome. *Nature*, **431**, 931-945. https://doi.org/10.1038/nature03001

[20] Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., *et al.* (2021) The Structure, Function and Evolution of a Complete Human Chromosome 8. *Nature*, **593**, 101-107. https://doi.org/10.1038/s41586-021-03420-7

[21] Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., *et al.* (2022) Semi-automated Assembly of High-Quality Diploid Human Reference Genomes. *Nature*, **611**, 519-531. https://doi.org/10.1038/s41586-022-05325-5

[22] Logsdon, G.A., Ebert, P., Audano, P.A., Loftus, M., Porubsky, D., Ebler, J., *et al.* (2025) Complex Genetic Variation in Nearly Complete Human Genomes. *Nature*, **644**, 430-441. https://doi.org/10.1038/s41586-025-09140-6

[23] Collins, F.S. and Mansoura, M.K. (2001) The Human Genome Project. *Cancer*, **91**, 221-225. https://doi.org/10.1002/1097-0142(20010101)91:1+<221::aid-cncr8>3.3.co;2-0

[24] NIH, National Human Genome Research Institute, Genetics vs. Genomics Fact Sheet. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

[25] Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S. and Schmidtke, J. (1985) An Estimate of Unique DNA Sequence Heterozygosity in the Human Genome. *Human Genetics*, **69**, 201-205. https://doi.org/10.1007/bf00293024

[26] Rudner, R., Karkas, J.D. and Chargaff, E. (1968) Separation of *B. subtilis* DNA into Complementary Strands. 3. Direct Analysis. *Proceedings of the National Academy of Sciences*, **60**, 921-922. https://doi.org/10.1073/pnas.60.3.921