Relating Digital Information, Thermodynamic Stability, and Classes of Functional Genes in E. coli

Dawit Nigatu, Werner Henkel
Transmission Systems Group
Jacobs University Bremen
{d.nigatu & w.henkel}@jacobs-university.de

Patrick Sobetzko, Georgi Muskhelishvili
Molecular Genetics Group
Jacobs University Bremen
{p.sobetzko & g.muskhelishvili}@jacobs-university.de

***Abstract*—Tremendous efforts have been made to analyze and discern the digital information content of the DNA ever since the introduction of the Watson-Crick model, later fueled by the availability of genomic data. However, there is also an analog type of information which is related to the physicochemical properties of the DNA, manifested in structural and topological variations of the chromosome. Hence, investigating the relationship between digital information contained in the sequence of bases and the analog parameters associated with it is very important to the general understanding of the coding structure in the DNA. In this paper, we represented analog information by thermodynamic stability and compare it with digital information using Shannon and Gibbs entropy measures on the complete genome sequence of the bacteria *Escherichia coli (E. coli)*. Furthermore, the link to the broader classes of functional gene groups (anabolic, catabolic, aerobic, and anaerobic) is examined. In most regions of the genome, the Shannon and Gibbs entropies are anti-correlated. Around the terminus, there is an almost perfect anti-correlation with high Shannon and low Gibbs entropies, meaning that the sequence is more random and at the same time less stable. The other core finding is the very high similarity in the profiles of entropies and the distribution of anabolic genes.**

***Index Terms*—Biological Sequence Analysis, DNA, Functional classes of Genes, Thermodynamic Stability**

I. Introduction

The information contained in the DNA is inscribed by the sequence of the four bases Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). The average information content of the genome can be measured using Shannon entropy [1]. So far, researchers have extensively applied this information theoretic measure for studying a wide variety of topics in molecular biology and bioinformatics, including DNA pattern recognition, gene prediction, sequence alignment, and comparative genomics [2]–[9]. We applied Shannon's entropy for identifying an underlying coding structure in a complete genome of an organism. However, we believe that solely looking at the base or codon composition in nucleotide sequences will not show the complete picture of the underlying coding structure in the DNA. In addition to the digital information, there is also an "analog" information present, which is due to physicochemical properties of the DNA [10] [11]. The three-dimensional analog information is a result of dynamic structural and topological variations of chromosomes (e.g. shape and stiffness) to facilitate and regulate activities such as gene expression, chromosome compaction, replication, and transcription [12]–[14]. Hence, looking into both the digital information in the nucleotide sequence and the analog coding counterpart jointly is very important.

Stacking between adjacent base pairs and pairing between complimentary bases determine the thermodynamic stability of the DNA [15] [16]. It is asserted that the relative stability of the DNA duplex structure relies on the identity and orientation of successive base steps [17] [18]. In bacteria, the physicochemical properties, including DNA thermodynamic stability, supercoiling, and mechanical stiffness, are dictated by interactions between neighboring bases and are central properties which, for example, determine the gene expression [12]. Since the stability of the DNA is a very decisive factor and due to the availability of thermodynamic parameters to describe DNA stability, such as Santalucia's unified nearest-neighbor (NN) thermodynamic stability parameters (free energies) of Watson-Crick base pairs in 1 M NaCl [19], we would associate analog information to be a measure of relative thermodynamic stability. However, we did not include sequence-independent effects of stability.

In this study, we will base our analysis and observations on the 4641652 bp long genome sequence of the *E. coli* K-12 MG1655 strain (accession number: [GenBank: U00096.3]). Shannon's block entropy is used to measure the digital information, whereas Gibbs' entropy is employed to measure the thermodynamic stability after the probability distribution is properly adapted to represent stability. To further relate the two forms of information with a functional meaning, we will also incorporate spatial distributions of the wide classes of anabolic, catabolic, aerobic, and anaerobic functional genes. By doing so, we hope to see connections between thermodynamic stability and digital information as well as functional meanings it might provide.

II. Methods

First, the genome sequence is rearranged to start at the origin (OriC) of replication. Then, the entropy of chunks of the DNA sequence is computed by scanning the complete genome with a sliding window. Within a window, all possible words of the given block size (N) are counted. To account for all adjacent base interactions, neighboring base pairs are considered. That is, if the nucleotide sequence is "AGCTAG" and the block size is 3 base pairs (bp), AGC, GCT, CTA, and

TAG are counted. In this paper, only a block size of three ($N = 3$) is considered (i.e. codons).

The Shannon entropy quantifies the average information content of the sequence from the distribution of symbols (words) of the source [20]. It is mathematically given as

$$H_N = -\sum_{i=1}^{64} P_s(i) \log P_s(i) \ , \tag{1}$$

where $P_s(i)$ is the probability (relative frequency) to observe the i^{th} codon. The Shannon entropy is maximal when all words occur at equal probabilities, and it is zero when one of the symbols occurs with probability one.

Ledwig Boltzmann was the first to give a statistical explanation of the physical (thermodynamic) entropy by relating it to the number of possible arrangements of molecules (microstates) belonging to a macrostate [21]. The celebrated formula reads

$$S_B = k_B \ln \Omega \ . \tag{2}$$

k_B is the Boltzmann constant which gives this entropy a thermodynamic unit of measure, $k_B = 1.38 \times 10^{-23} J/K$, and Ω is the number of accessible microstates. Boltzmann's entropy is defined for a system based on a microcanonical ensemble in which the macrostate is of a fixed number of particles, volume, and energy. All states are accessed equally likely with the same energy [22]. Later, Gibbs devised a generic entropy definition over the more general probability distribution of the possible states (canonical ensemble). The Gibbs entropy is defined as

$$S_G = -k_B \sum_i P_G(i) \ln P_G(i) \ , \tag{3}$$

where the sum is over all microstates and $P_G(i)$ is the probability that the molecule is in the i^{th} state. It can easily be seen that for a uniform distribution of states, the Gibbs entropy reduces to the Boltzmann entropy.

Gibbs' entropy has a similar form as Shannon's entropy except for the Boltzmann constant. Nevertheless, unlike the Shannon case where the probability P_s is defined according to the frequency of occurrence, we will associate the probability distribution with thermodynamic stability quantified by the nearest-neighbor free energy parameters. We used Santalucia's unified free energy parameters for di-nucleotide steps at $37^0 C$ in [19], presented here in Table 1. For block sizes greater than two, the energies are computed by adding the involved di-nucleotides. For instance, if the block size is three and the sequence is AGC, the energies of AG and GC will be added. This way, we have a list of codons with their corresponding energies, providing 64 energy states denoted by $E(i)$. Assuming a random process behind the construction of the DNA, with a certain probability, one would obtain molecules with certain energies. If there are n_i codons in the i^{th} energy state, we assumed that the probability for having a certain energy state follows the Boltzmann distribution given by

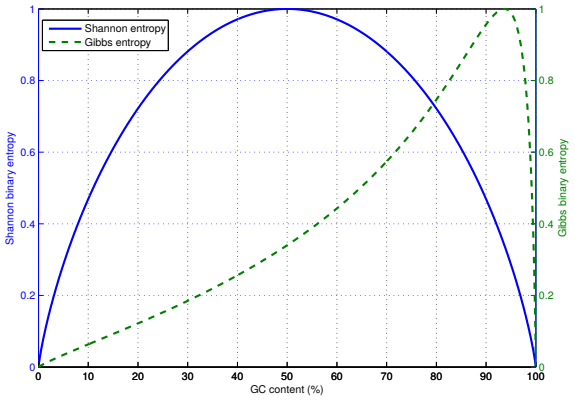

Fig. 1. Shannon and Gibbs entropies for binary input sequence

$$P_G(i) = \frac{n_i e^{-\frac{E(i)}{k_B T}}}{\sum_{j=1}^{64} n_j e^{-\frac{E(j)}{k_B T}}} \ . \tag{4}$$

T is the temperature in Kelvin. Although we are aware that the Boltzmann distribution gives the most probable distribution of energy (the one pertaining to the equilibrium state) for states having a random distribution of energies (e.g. ideal gas), which is not the case here, we just used it to have a representation of stability (energy) in an entropy-like expression.

To see how the Gibbs entropy captures the stability, let us consider the binary case where there are only two possible entries, AT and GC. One representing the AT-richness and the other GC-richness. If the probability of GC is p, the probability of AT will be $1-p$. We use the energies of AT and GC from the table and compute both entropies by changing the GC richness p from 0 % to 100 %, stepwise. Ignoring the Boltzmann constant and using the same logarithmic base, the result is shown in Fig. 1. The binary Shannon entropy function is symmetric with the maximum at 50 %. It tells us how random the sequence is. By comparing it with the maximum value, we can tell how diverse the sequence is, but it does not distinguish between AT and GC. However, the Gibbs entropy curve is uniformly related to the GC content (except for extremely large values of p, which is not feasible because the GC content of organisms typically cannot be greater than 80 %). The higher the Gibbs entropy, the higher the GC content. Hence, it measures stability.

The functional gene groups were taken from the Gene Ontology (GO) branches provided by the RegulonDB database. Anabolic genes: biosynthesis of macromolecules (GID000000120); catabolic genes: degradation of macromolecules (GID000000057); aerobic genes: aerobic respiration (GID000000068); anaerobic genes: anaerobic respiration (GID000000069). In 500 kbp sliding windows of 4 kbp shift, the corresponding functional groups where counted. The window size was chosen to have a significant number of genes to obtain a smooth curve.

TABLE I
UNIFIED NEAREST-NEIGHBOR FREE ENERGY PARAMETERS OF WATSON-CRICK BASE PAIRS IN 1 M NACL AT 37^0C [19].

Sequence	AA|TT	AT	TA	CA|TG	GT|AC	CT|AG	GA|TC	CG	GC	GG|CC
$\Delta G(Kcal/mol)$	-1.00	-0.88	-0.58	-1.45	-1.44	-1.28	-1.30	-2.17	-2.24	-1.84

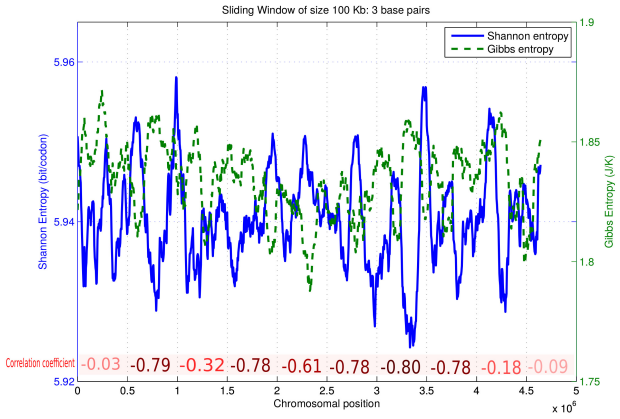

Fig. 2. Shannon and Gibbs entropies per codon for window size of 100 kb

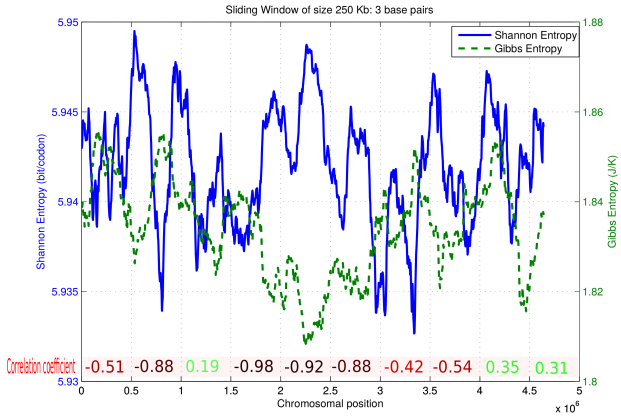

Fig. 3. Shannon and Gibbs entropies per codon for window size of 250 kb

III. RESULTS AND DISCUSSION

Our first aim was to compare the "analog" information, quantifying relative stability and measured with the Gibbs entropy (applying Boltzmann statistics to convert the stacking or melting energies to probabilities), with the digital Shannon information. To do so, a sliding window is shifted 4 kb at a time along the complete genome starting from the origin (OriC) as the center of the first window. To support our qualitative statements of comparisons, region-wise cross-correlation coefficients are incorporated in the figures. The Shannon and Gibbs entropies are plotted together for window sizes of 100 kb, 250 kb, and 500 kb in figures 2 to 4. Since the nucleotide sequence is rearranged to start at the origin of replication (OriC), the terminus region will be exactly in the middle. This is also evidently visible from the shape of Gibbs entropy curve in which the lowest point is around the terminus, attributed to the AT-richness. Smaller windows lead to high fluctuations and are not easy to compare. Likewise, a very large window will hinder the visibility of the differences as a result of the smoothing effect it creates. However, the general shapes of the curves are not affected by a change in window size. The positions of the troughs and crests are preserved.

The two entropies are mostly anti-correlated, with a stronger magnitude around the terminus. The terminus region is characterized by high Shannon entropy and low Gibbs entropy, more random and less stable. This means, the codon composition of the sequence has become slightly more balanced, which is due to an increase in AT-rich codons. Similarly, there are also positions where the Shannon entropy is relatively low and the Gibbs entropy is higher (e.g. in Fig. 3 around positions 0.8 Mbp and 3.4 Mbp) which means a codon bias towards being more GC-rich. In general, the interpretation for a block size of 3 bp should go as follows. If both entropies increase,

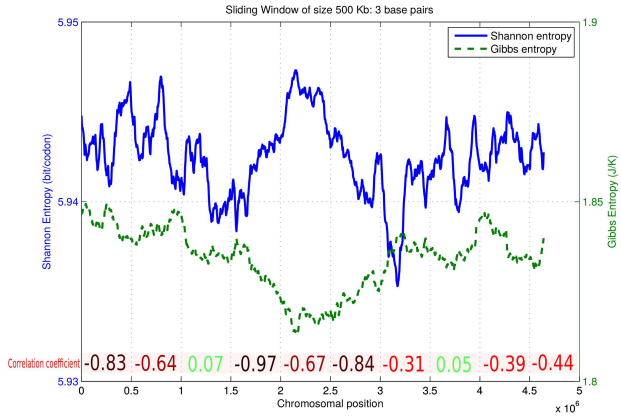

Fig. 4. Shannon and Gibbs entropies per codon for window size of 500 kb

it means that both the GC content and the randomness has increased, the sequence is stabilized by the usage of more GC-rich codons. However, if there is a decrease in Gibbs entropy while Shannon entropy is higher, the sequence has become less stable (AT-rich) and more random as a result of an increase in AT-rich codons usage.

To see what functional meaning resides in the corresponding chromosomal regions, we further compared the Shannon and Gibbs entropies with the distribution of the four classes of functional genes, namely anabolic, catabolic, aerobic, and anaerobic. We used a 500 kb window and counted the number of genes of the corresponding functional group. The distribution of the four classes of genes are plotted along with the Gibbs entropy. The results are presented in Fig. 5. Interestingly, from the figure, we observe that the shape of Gibbs entropy and the distributions of anabolic genes are

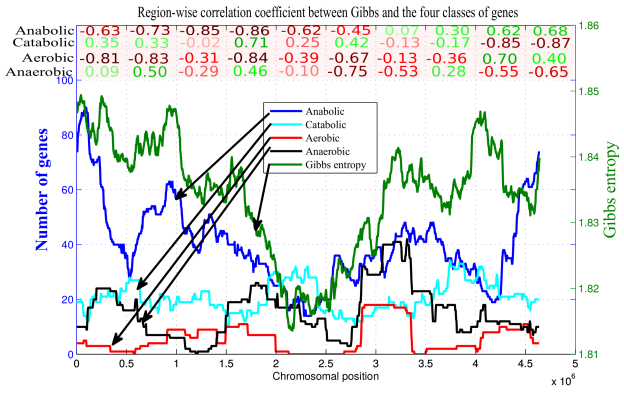

Fig. 5. Distribution of functional classes of genes and Gibbs entropy in 500 kb sliding window

strongly related. In a nutshell, the trends for the Gibbs entropy and the number of anabolic genes are decreasing as we move from the origin to the terminus along the chromosome. From the region-wise correlation values, it can be seen that a high cross correlation exists, especially in the right replichore. The global cross-correlation coefficient between Gibbs entropy and the presence of anabolic genes is around 0.64, which is also a high correlation considering the length of the genome. The distribution of the aerobic genes has also the same increasing and decreasing patterns as the Gibbs entropy, except for the quantization effects resulting from a very low number of genes. The dependency of the catabolic and anaerobic genes with the entropies is not so uniform and obvious to see as for anabolic and aerobic genes. However, there are relationships in distinct regions of the chromosome. Overall, the plots yield qualitative relations between digital and analog quantities or properties and gene functions at specific sites on the chromosome.

IV. Conclusion

Analyzing the thermodynamic stability and digital informations jointly not only provides an additional angle to interpreting and understanding the genome sequence but also provides a way to incorporate the structural and functional implications as well. We have shown the connection between the digital information of the sequence, the relative thermodynamic stability, and the functional meaning of genes. The region-wise cross-correlation coefficients show that Shannon and Gibbs entropies are mostly anti-correlated in the *E. coli* genome. Especially, the two entropies are almost exactly opposite around the terminus, which is a justification of the low stability and more uniform distribution of codons. The other core finding is the relation between the distribution of anabolic and aerobic genes to the Gibbs entropy. The observed patterns are very similar, implying the clear connection between gene types and stability and, due to the correlation between entropies, also to statistical properties, i.e., the information content.

V. Acknowledgment

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948. [Online]. Available: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[2] S. Akhter, B. A. Bailey, P. Salamon, R. K. Aziz, and R. A. Edwards, "Applying shannon's information theory to bacterial and phage genomes and metagenomes," *Scientific reports*, vol. 3, 2013.

[3] C.-H. Chang, L.-C. Hsieh, T.-Y. Chen, H.-D. Chen, L. Luo, and H.-C. Lee, "Shannon information in complete genomes," *Journal of bioinformatics and computational biology*, vol. 3, no. 03, pp. 587–608, 2005.

[4] T. D. Schneider, "A brief review of molecular information theory," *Nano communication networks*, vol. 1, no. 3, pp. 173–180, 2010.

[5] T. D. Schneider and J. Spouge, "Information content of individual genetic sequences," *J. Theor. Biol.*, vol. 189, pp. 427–441.

[6] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, vol. 188, pp. 415–431, 1986.

[7] R. Roman-Roldan, P. Bernaola-Galvan, and J. Oliver, "Application of information theory to dna sequence analysis: a review," *Pattern recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[8] E. Capriotti, P. Fariselli, I. Rossi, and R. Casadio, "A shannon entropy-based filter detects high-quality profile–profile alignments in searches for remote homologues," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 54, no. 2, pp. 351–360, 2004.

[9] J. Hagenauer, Z. Dawy, B. Gobel, P. Hanus, and J. Mueller, "Genomic analysis using methods from information theory," in *Information Theory Workshop, 2004. IEEE.* IEEE, 2004, pp. 55–59.

[10] G. Muskhelishvili and A. Travers, "Integration of syntactic and semantic properties of the dna code reveals chromosomes as thermodynamic machines converting energy into information," *Cellular and Molecular Life Sciences*, vol. 70, no. 23, pp. 4555–4567, 2013.

[11] A. Travers, G. Muskhelishvili, and J. Thompson, "DNA information: from digital code to analogue structure," *Philos Transact A Math Phys Eng Sci*, vol. 370, no. 1969, pp. 2960–86, 2012.

[12] P. Sobetzko, M. Glinkowska, A. Travers, and G. Muskhelishvili, "DNA thermodynamic stability and supercoil dynamics determine the gene expression program during the bacterial growth cycle," *Mol BioSyst*, vol. 9, no. 7, pp. 1643–1651, 2013.

[13] A. Travers and G. Muskhelishvili, "Dna thermodynamics shape chromosome organisation and topology," *Biochem Soc Trans*, vol. 41, pp. 548–553, 2013.

[14] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt, "Analog regulation of metabolic demand," *BMC systems biology*, vol. 5, no. 1, p. 40, 2011.

[15] E. Protozanova, P. Yakovchuk, and M. D. Frank-Kamenetskii, "Stacked–unstacked equilibrium at the nick site of dna," *Journal of molecular biology*, vol. 342, no. 3, pp. 775–785, 2004.

[16] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the dna double helix," *Nucleic acids research*, vol. 34, no. 2, pp. 564–574, 2006.

[17] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky, "Predicting dna duplex stability from the base sequence," *Proceedings of the National Academy of Sciences*, vol. 83, no. 11, pp. 3746–3750, 1986.

[18] J. SantaLucia Jr and D. Hicks, "The thermodynamics of dna structural motifs," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 415–440, 2004.

[19] J. Santalucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 4, pp. 1460–1465, 1998.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[21] R. Bowley and M. Sánchez, *Introductory statistical mechanics*. Oxford: Clarendon Press, 1999.

[22] F. Reif, *Fundamentals of Statistical and Thermal Physics*, international student edition ed. New York: McGraw-Hill, 1985.