Entropy in DNA

Lucas Blakeslee[1]

[1] *Institute for Computing in Research. Santa Fe, New Mexico, United States of America*

July 23, 2021

The entropy of a genomic sequence can provide a variety of insights into its nature, namely in that it can be an indicator of whether a region of DNA is expressed. This paper attempts to lay out the foundations for a new method of calculating entropy by bringing in information taken from the sequence's biological class. A normalized function was described that can identify the class of a sequence via subsequence comparison, which down the line can be used to create a definition for entropy specific to DNA.

1 Background

Entropy as a measure of information content was described by Shannon (1948), whereupon he contributed immensely to founding the modern field of information theory. Where H is the entropy, and p_i is the probability of a system being in state i (which can be thought of as the probability of i occurring), Shannon entropy is defined as:

$$H(P) = -\sum_{i=1} p(i) \log p(i) \tag{1}$$

Dividing the Shannon entropy by the information length can serve as normalization, resulting in what is called metric entropy [1].

An additional measure of entropy is topological entropy, originally designed as a descriptor of topological dynamical systems, but implemented for strings of finite length by Koslicki (2011). Koslicki's definition is as follows:

If w is a finite sequence of length $|w|$, and n is the unique integer such that: $4^n + n - 1 \leq |w| < 4^{n-1} + (n+1) - 1$

$$H_{top}(w) = \frac{\log_4(p_{w^{4^n+n-1}}(n))}{n} \tag{2}$$

The most important conceptual difference between Shannon's and Koslicki's entropies is that the application of topological entropy ignores the frequency with which subsequences occur.

The concept of entropy has been applied to a variety of physical systems, but what about biological ones? The question of entropy as it relates to biology was essentially first explored in Erwin Schrödinger's 1944 book *What is Life?*, wherein he asked fundamental questions regarding how life requires *negentropy* (in a later edition corrected by the author to free energy) in order to keep from decaying, and maintaining the decrease of entropy that evolution brings.

[1] It is worth noting that there are multiple definitions of metric entropy, and the term often refers to entropy of a metric space

Entropy as a measure of the complexity held within *DNA* sequences has been described in the Shannonian sense by Schmitt & Herzel (1997). The idea of block entropies is crucial for understanding the entropy of a DNA sequence. For a given alphabet $\aleph$ (here {A, G, T, C}), entropy will be low in a sequence S when symbols and/or sub-sequences repeat. Classic Shannon entropy provides insight into the probabilities with which symbols occur, however, it does not provide insight into the relationships between different symbols. For that, *block entropies* are required. The block entropy of a sequence can be defined as:

$$H_n = \sum_i p_i^{(n)} \log p_i^{(n)} \tag{3}$$

Where $P_i^{(n)}$ are the probabilities of the combinations of n symbols.

The entropy of a DNA sequence can also bear insight into intron and exon regions (regions of DNA that aren't expressed and those that are), as intron regions aren't subject to the same evolutionary pressures to which exon regions are exposed, thus they would be expected to have a higher entropy (Koslicki, 2011).

2 Motivation

Here, I wondered whether a calculation of entropy for biological sequences could be defined given information about the class of the sequence. The idea here was that if knowledge was already had about common subsequences in a given sequence's class, an "entropy reduction" could be created to subtract from the raw entropy (which could be Shannonian or topological).

3 Methods

The first step in this process was being able to classify sequences into a biological class based on the subsequences (equivalent to blocks in block entropy) within the class of a genome. For proof of concept, two classes were used here: epsilonproteobacteria and gammaproteobacteria. For many members of each class, all subsequences up to length 15 within the bacterial genomes were counted, and the number of occurrences of each subsequence were averaged for each class.

It was hypothesized that the Poisson distribution, a probability distribution that can estimate the number of occurrences of an event within a given period, could be used to predict the probabilities that an unknown sequence belongs to a given class based on the subsequences in the unknown sequences and the expected number of subsequences per class.

The classic poisson distribution is:

$$P(\lambda, k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4}$$

From which the following was derived:

$$\ell(k_i,^k \lambda_i) = \sum k_i \ln^t \lambda_i + c \tag{5}$$

where ℓ is the log probability.

To put this into the terms of subsequences and classes:

$$\ell(C_i, C_i\_expected\_T) = \sum C_i ln(C_i\_expected\_T) + K \tag{6}$$

Where C_i is the count of i^{th} subsequence and T is the type of organism. K is a constant unrelated to $C_i\_expected\_T$.

The above was implemented, and the following plot was produced:

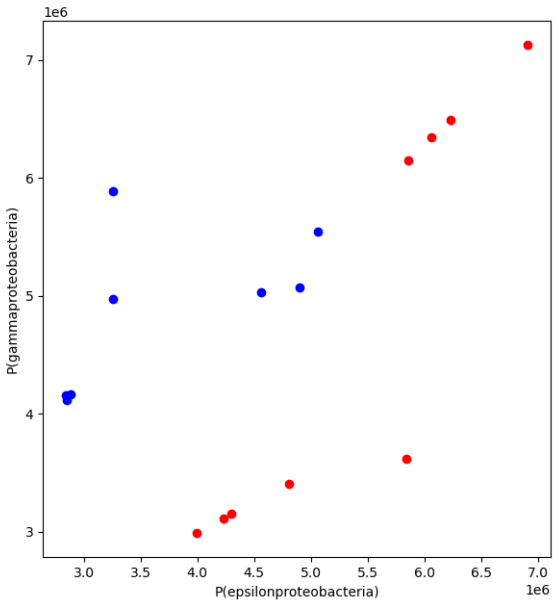

Figure 1

Where blue dots are epsilonproteobacteria sequences, red dots are gammaproteobacteria sequences, and the (x, y) axes are the calculated probabilities for a sequence falling into the epsilonproteobacteria and gammaproteobacteria classes respectively.

There is not clear separation, and the results of the function turned out not to be very meaningful for identifying species. One hypothesis as to why this might be has to do with microsatellites, which are blocks of DNA that repeat thousands to tens of thousands of times within a genome. Microsatellites are repeating sequences composed of small repeating blocks, such as $GCGCGCGC$, or $ATCATCATCATC$. Microsatellites may well not be Poisson random, and additionally do not seem to provide any special insight into what class a species falls into.

With that in mind, the following function f given the counts in an unknown sequence C_i, and the expected counts averaged from a class E_i was defined:

$$f = \frac{\sum(\log(C_i)\log(E_i))}{\sqrt{\sum(\log(C_i)^2)\sum(\log(E_i)^2))}} \tag{7}$$

Where the denominator serves as a normalizing constant.

4 Results

The output of equation 7 yielded much more clear separation between members of different classes than can be seen in Figure 1.

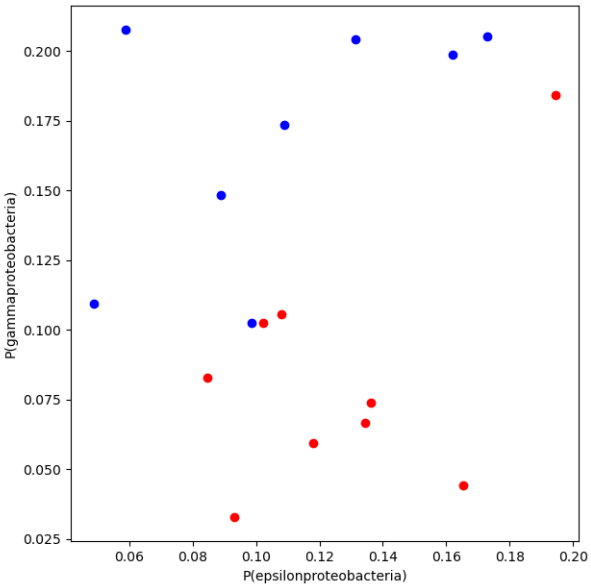

Figure 2

A more-or-less diagonal separation can be seen between the gamma- and epsilonproteobacteria based on their calculated probabilities with a given class.

Unfortunately, due to time constraints, the project had to be left here, but I plan to continue work on this project in a few areas.

4.1 Future Work

Going forward, crafting a definition for an entropy reduction to be subtracted from raw entropy is of prime importance. Performing these calculations on organisms of other classes (namely on eukaryotes and species with larger genomes and known intron regions) should yield further insight. Ultimately, I would like to map the genomes of organisms in an attempt to identify intron and exon regions, and see how the accuracy of intron/exon identification compares between that method, Shannon, and topological entropies.

5 Acknowledgments

I would sincerely like to thank my mentor, David Palmer of the Los Alamos National Laboratory, for the truly invaluable guidance and help he has provided me throughout this process.

Additionally, I would like to thank the Institute for Computing in Research for providing me with the opportunity to perform this research.

6 References

Koslicki, D. (2011). Topological entropy of DNA sequences. *Bioinformatics*, 27(8), 1061–1067. https://doi.org/10.1093/bioinformatics/btr077

Schmitt, A. O., & Herzel, H. (1997). Estimating the Entropy of DNA Sequences. *Journal of Theoretical Biology*, 188(3), 369–377. https://doi.org/10.1006/jtbi.1997.0493

Schrödinger, E. (1944). What is Life? *Cambridge University Press*

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. *Urbana: University of Illinois Press*.