BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 8 2011, pages 1061–1067
doi:10.1093/bioinformatics/btr077

Sequence analysis

Advance Access publication February 10, 2011

Topological entropy of DNA sequences

David Koslicki
Department of Mathematics, Pennsylvania State University, State College, PA 16801, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Topological entropy has been one of the most difficult to implement of all the entropy-theoretic notions. This is primarily due to finite sample effects and high-dimensionality problems. In particular, topological entropy has been implemented in previous literature to conclude that entropy of exons is higher than of introns, thus implying that exons are more 'random' than introns.

Results: We define a new approximation to topological entropy free from the aforementioned difficulties. We compute its expected value and apply this definition to the intron and exon regions of the human genome to observe that as expected, the entropy of introns are significantly higher than that of exons. We also find that introns are less random than expected: their entropy is lower than the computed expected value. We also observe the perplexing phenomena that introns on chromosome Y have atypically low and bimodal entropy, possibly corresponding to random sequences (high entropy) and sequences that posses hidden structure or function (low entropy).

Availability: A Mathematica implementation is available at http://www.math.psu.edu/koslicki/entropy.nb

Contact: koslicki@math.psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 29, 2010; revised on January 17, 2011; accepted on February 4, 2011

1 INTRODUCTION

Entropy, as a measure of information content and complexity, was first introduced by Shannon (1948). Since then entropy has taken on many forms, namely topological, metric (due to Shannon), Kolmogorov–Sinai and Rènyi entropy. These entropies were defined for the purpose of classifying a system via some measure of complexity or simplicity. These definitions of entropy have been applied to DNA sequences with varying levels of success. Topological entropy, in particular, is infrequently used due to high-dimensionality problems and finite sample effects. These issues stem from the fact that the mathematical concept of topological entropy was introduced to study *infinite* length sequences. It is universally recognized that the most difficult issue in implementing entropy is the convergence problem due to finite sample effects (Vinga and Almeida, 2004; Kirillova, 2000). A few different approaches to circumvent these problems with topological entropy and adapt it to *finite* length sequences have been attempted before. For example, in Troyanskaya *et al.* (2002), linguistic complexity (the fraction of total subwords to total possible subwords) is utilized to circumvent finite sample problems. This leads to the observation that the complexity/randomness of intron regions is *lower* than the complexity/randomness of exon regions. However in Colosimo and de Luca (2000), it is found that the complexity of randomly produced sequences is *higher* than that of DNA sequences, a result one would expect given the commonly held notion that intron regions of DNA are free from selective pressure and so evolve more randomly than do exon regions.

Also, little has been done in the way of mathematically analyzing other finitary implementations of entropy due to most previous implementations using an entire function instead of a single value to represent entropy (thus the expected value would be very difficult to calculate).

In this article, we focus on topological entropy, introducing a new definition that has all the desired properties of an entropy and still retains connections to information theory. This approximation, as opposed to previous implementations, is a *single* number as opposed to an entire function, thus greatly speeding up the calculation time and removing high-dimensionality problems while allowing more mathematical analysis. This definition will allow the comparison of entropies of sequences of differing length, a property no other implementation of topological entropy has been able to incorporate. We will also calculate the expected value of the topological entropy to precisely draw out the connections between topological entropy and information content. We will then apply this definition to the human genome to observe that the entropy of intron regions is in fact lower than that of exon regions as one would expect.

2 METHODS

2.1 Definitions and preliminaries

We restrict our attention to the alphabet $\mathcal{A}=\{A,C,T,G\}$. For a finite sequence w over the alphabet $\mathcal{A}$, we use $|w|$ to denote the length of w. Of primary importance in the study of topological entropy is the complexity function of a sequence w (finite or infinite) formed over the alphabet $\mathcal{A}$.

DEFINITION 1 (complexity function). *For a given sequence w, the complexity function* $p_w:\mathbb{N}\to\mathbb{N}$ *is defined as*

$$p_w(n)=|\{u:|u|=n \text{ and u appears as a subword of } w\}|$$

That is, $p_w(n)$ represents the number of different n-length subwords (overlaps allowed) that appear in w.

Now the traditional definition of topological entropy of an *infinite* word w is the asymptotic exponential growth rate of the number of different subwords:

DEFINITION 2. *For an infinite sequence w formed over the alphabet* $\mathcal{A}$*, the topological entropy is defined as*

$$\lim_{n\to\infty}\frac{\log_4 p_w(n)}{n}$$

Due to the limit in the above definition, it is easily observed that this definition will always lead to an answer of zero if applied directly to finite length sequences. This is due to the fact that the complexity function of

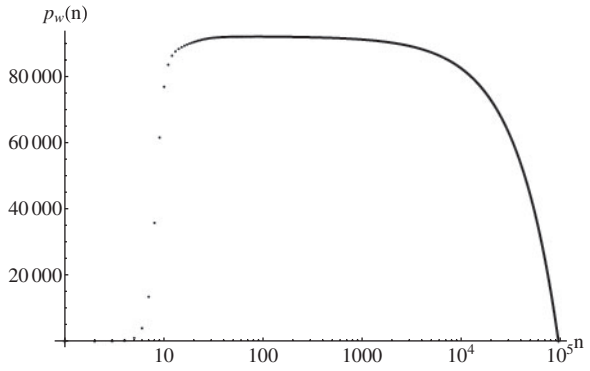

Fig. 1. Log-linear plot of the complexity function of the gene *ACSL4*.

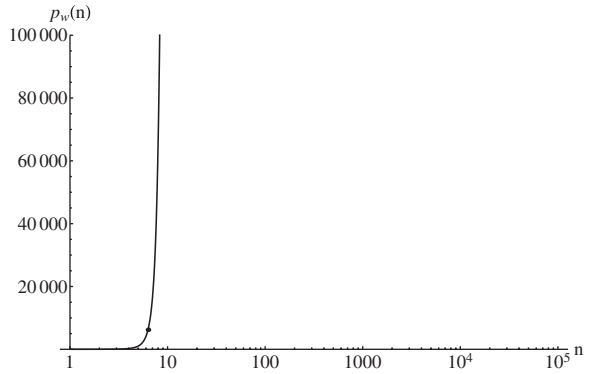

Fig. 2. Log-linear plot of the complexity function of a random infinite sequence.

infinite length sequences is non-decreasing, while of finite length sequences it is eventually zero. We include in Figures 1 and 2 a log-linear plot of the complexity functions for the gene *ACSL4* found on ChrX:108906440-108976621 (hg19) as well as for an infinite string generated by a Markov chain on four states with equal transition probabilities.

The graph of the complexity function of the gene found in Figure 1 is entirely typical of the graph of a complexity function for a finite sequence. The precise description of the shape of the complexity function can be found in the nice summary by Colosimo and de Luca (2000). In essence, on three disjoint intervals the complexity function $p_w(n)$ is strictly increasing, then non-decreasing,and then decreasing at a slope of -1.

Now for a finite sequence w, we desire that an approximation of topological entropy $H_{\text{top}}(w)$ should have the following properties:

(1) $0 \leq H_{\text{top}}(w) \leq 1$

(2) $H_{\text{top}}(w) \approx 0$ if and only if w is highly repetitive (contains few subwords)

(3) $H_{\text{top}}(w) \approx 1$ if and only if w is highly complex (contains many subwords)

(4) For different length sequences v, w, $H_{\text{top}}(w)$ and $H_{\text{top}}(v)$ should be comparable

It should be noted that item 4 on this list is of utmost importance when implementing topological entropy. It is very important to normalize with respect to length since otherwise when counting the number of subwords, longer sequences will appear artificially more complex simply due to the fact that since the sequence is longer, there are more chances for subwords to show up. This explains the 'linear correlation' between sequence length and the implementations of topological entropy used in Karamanos *et al.* (2006) and Kirillova (2000). This also hints at the incomparability of the notions of entropy contained in Karamanos *et al.* (2006), Colosimo and de Luca (2000), Kirillova (2000), and Schmitt and Herzel (1997).

Since topological entropy should give an approximate asymptotic exponential growth rate of the number of subwords, the only pertinent values of $p_w(n)$ are those in the interval where $p_w(n)$ is strictly increasing. With an alphabet of four letters, the maximal such value occurs at the smallest n such that $|w| < 4^{n+1} + (n+1) - 1$.

We define the approximation to topological entropy as follows:

DEFINITION 3 (topological entropy). *Let w be a finite sequence of length $|w|$, let n be the unique integer such that*

$$4^n + n - 1 \leq |w| < 4^{n+1} + (n+1) - 1$$

Then for $w_1^{4^n+n-1}$ the first $4^n + n - 1$ letters of w,

$$H_{\text{top}}(w) := \frac{\log_4(p_{w_1^{4^n+n-1}}(n))}{n}$$

The reason for concatenating w to the first $4^n + n - 1$ letters is due to the following two facts whose proofs can be found in the Supplementary Material.

LEMMA 1. *A sequence w over the alphabet $\{A, C, T, G\}$ of length $4^n + n - 1$ can contain at most 4^n subwords of length n. Conversely, if a word w is to have 4^n subwords, it must have length at least $4^n + n - 1$.*

Thus, if we had taken an integer $m > n$ in the above definitions and instead utilized $\frac{\log_4(p_w(m))}{m}$, w would not be long enough to contain all different possible subwords.

LEMMA 2. *Say a sequence w has length $4^n + n - 1$ for some integer n, if w contains all possible subwords of length n formed on the alphabet $\{A, C, T, G\}$, then $H_{\text{top}}(w) = 1$.*

Thus, if a sequence of length $4^n + n - 1$ is 'as random as possible' (i.e. contains every possible subword), its topological entropy is 1, just as we would expect in the infinite sequence case. Similarly, if w is 'as non-random as possible', that is, if w is simply the repetition of a single letter $4^n + n - 1$ times, then $H_{\text{top}}(w) = 0$.

Furthermore, if we had not used truncation in definition 3, then for a sequence v such that $|v| > |w|$, the topological entropy of v would on average be artificially higher due to v being a longer sequence and thus has more opportunity for the appearance of subwords. By truncating, we have allowed sequences of different lengths to have comparable topological entropies.

This definition of topological entropy serves as a measure of the randomness of a sequence: the higher the entropy, the more random the sequence. The justification for this finite implementation giving an approximate characterization of randomness is given in Ornstein and Weiss (2007) in which it is shown that functions of entropy are the only finitely observable invariants of a process.

2.2 Expected value

While topological entropy has been well studied for infinite sequences, very little has been done by the way of mathematically analyzing topological entropy for finite sequences. This lack of analysis is most likely due to topological entropy as in the literature (Crochemore and Vèrin 1999; Kirillova 2000; Schmitt and Herzel 1997) being considered not as a single number to be associated to a DNA sequence, but rather the entire function $\frac{\log_4 p_w(n)}{n}$ is considered for *every* n. This approach turns topological entropy (which should be just a single number associated to a DNA sequences) into a very high-dimensional problem. In fact, as many dimensions as is the length of the DNA sequence under consideration. Our approximation (Definition 3)

does in fact associate just a single number (instead of an entire function) to a sequence, and so is much more analytically tractable.

We now utilize the results of Gheorghiciuc and Ward (2007) to compute the expected value of topological entropy. This will assist in determining what constitutes 'high' or 'low' entropy. First, we calculate the expected value of the complexity function $p_w(n)$. As is commonly assumed (Hasegawa *et al.*, 1985; Jukes and Cantor, 1969), we now assume that DNA sequences evolve in the following way: each state in a Markov fashion independent of neighboring states. We do not assume a single model of molecular evolution, but rather just assume that there is some set of probabilities $\{\pi_A, \pi_C, \pi_T, \pi_G\}$ such that the probability of appearance of a sequence w is given by the following: for n_A the number of occurrences of the letter A in w, n_C the number of occurrences of the letter C in w, etc., the probability of the sequence w appearing is given by:

$$\mathbb{P}(w) = \pi_A^{n_A} \pi_C^{n_C} \pi_T^{n_T} \pi_G^{n_G}$$

This assumption regarding the probability of appearance of a DNA sequence is used only to procure a distribution against which we may calculate the expected number of subwords. The actual calculation of topological entropy as in Definition 3 does not make any such assumption about the probability of appearance.

THEOREM 1 (expected value of the complexity function). *The expected value of the complexity function $p_w(n)$ taken over sequences of length $|w| = n+k-1$ is given by*

$$\mathbb{E}[p_w(n)] = 4^n - \sum_w (1 - \mathbb{P}(w))^k + \mathcal{O}(k^{-\epsilon}\mu^n) \tag{1}$$

where the summation is over all sequences w of length n, $0 < \epsilon < 1$, and $\mu < 1$ [these are explicitly computed constants based on the π_i defined above, see (Gheorghiciuc and Ward, 2007)].

PROOF. See (Gheorghiciuc and Ward, 2007).

This theorem has a particularly nice reduction [Gheorghiciuc and Ward (2007), Corollary 2.2] when one assumes that the probability of appearance of each subletter is the same (equivalent to the the expected value being computed with a uniform distribution on the set of all sequences of a certain length). While clearly there *is* a mononucleotide bias for different genomic regions and DNA sequences do not occur uniformly randomly, we do assume equal probability of appearance of each nucleotide as then the calculation of the expected number of subwords reduces in computational complexity from exponential to linear in the length of the sequence.

It is a straightforward calculation to combine Corollary 2.2 in Gheorghiciuc and Ward (2007) with Definition 3 and compute the constants ϵ and μ. Doing so, we obtain the following expected value for the topological entropy.

THEOREM 2 (expected value of topological entropy). *The expected value of topological entropy taken over sequences of length $|w| = 4^n + n - 1$ is given by*

$$\mathbb{E}[H_{\text{top}}] = \frac{\log_4(4^n - 4^n(1 - 1/4^n)^{4^n} + \mathcal{O}((\frac{1}{\sqrt{2}}))^n)}{n} \tag{2}$$

We now present in Table 1 the calculated estimation of the expected value of H_{top} using the above formula. Keep in mind that the convergence of this calculation to the actual expected value is exponentially quick as n increases (and so also the length of the sequence). We thus ignore the $\mathcal{O}((\frac{1}{\sqrt{2}}))^n$ term in the following calculation.

For comparison's sake, we present in Table 2 the sampled expected values for $n = 1, \ldots, 9$ along with sampled SDs (the calculation were made by explicitly computing the topological entropy of uniformly randomly selected sequences).

Summarizing this table, the topological entropy of randomly selected sequences is tightly centered around the expected value which itself is close to one. Furthermore, the distribution of topological entropy is very close to a normal distribution as can be observed from the histogram of topological entropy for sequences of length $4^9 + 9 - 1$ included in Figure 3. The skewness and kurtosis are 0.0001996 and 2.99642, respectively.

Table 1. Calculated expected value of topological entropy

n	$4^n + n - 1$	Calculated expected value of H_{top}
1	4	0.725606
2	17	0.841242
3	66	0.890810
4	249	0.917489
5	1028	0.933868
6	4101	0.944865
7	16390	0.952736
8	65543	0.958642
9	262152	0.963237
10	1048585	0.966914
11	4194315	0.969921
12	16777227	0.972428

Table 2. Sampled expected value and SD of topological entropy

n	$4^n + n - 1$	Sampled expected value of H_{top}	Sampled SD	Sample size
1	4	0.703583	0.184798	256
2	17	0.838956	0.0508640	300 000
3	66	0.890576	0.0176785	300 000
4	249	0.917457	0.00674325	300 000
5	1028	0.933869	0.0027160	300 000
6	4101	0.944861	0.00113176	300 000
7	16390	0.952733	0.000486 368	300 000
8	65543	0.958642	0.000212283	300 000
9	262152	0.963237	0.0000944814	300 000

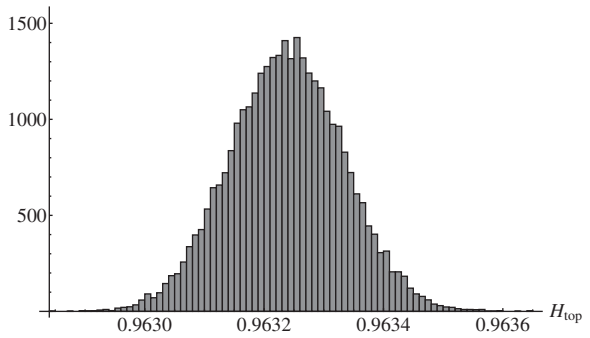

Fig. 3. Histogram of topological entropy of randomly selected sequences of length $4^9 + 9 - 1 = 262152$.

3 ALGORITHM

An implementation of this approximation to topological entropy is available at: http://www.math.psu.edu/koslicki/entropy.nb

We mention a few notes regarding this estimation of topological entropy. First, if a sequence w in consideration has a length such that for some n, $4^n + n - 1 \ll |w| < 4^{n+1} + n$, it will be more accurate to use a sliding window to compute the topological entropy. For

example, if $|w| = 16000$, we would normally truncate this sequence to the first 4101 letters. This might misrepresent the actually topological entropy of the sequence. Accordingly, we could instead compute the average of the topological entropy of the following sequences (where w_n^m means the subsequence of w consisting of the n-th to m-th letters of w):

$$w_1^{4101}, w_2^{4102}, w_3^{4103}, \ldots, w_{11899}^{16000}$$

This is computationally intensive, so for longer sequences, one might instead choose to take non-overlapping windows. That is, finding the average of the topological entropy of the sequences

$$w_1^{4101}, w_{4102}^{8203}, w_{8204}^{12305}, \ldots$$

The above web site includes serial and parallel versions of the algorithm. The fastest version utilizes Nvidia CUDA GPU computing, has complexity $\mathcal{O}(n)$ for a sequence of length n and takes an average of 5.2 s to evaluate on a DNA sequence of length 16 777 227 when using an Intel i7-950 3.6 GHz CPU and an Nvidia GTX 460 GPU.

3.1 Comparison to traditional measures of complexity

Other measures of DNA sequence complexity similar to this approximation of topological entropy include: previous implementations of topological entropy (Kirillova, 2000), special factors (Colosimo and de Luca, 2000), Shannon's metric entropy (Farach *et al.*, 1995; Kirillova, 2000), Rènyi continuous entropy (Rènyi, 1961; Vinga and Almeida, 2004) and linguistic complexity (LC) (Gabrielian and Bolshoy, 1999; Troyanskaya *et al.*, 2002).

The implementation of topological entropy in Kirillova (2000) does not produce a single number representing entropy, but rather an entire sequence of values. Thus, while the implementation of Kirillova (2000) does distinguish between artificial and actual DNA sequences, Kirillova notes that the implementation is hampered by high-dimensionality and finiteness problems.

In Colosimo and de Luca (2000), it is noted that the special factors approach does not differentiate between introns and exons.

Note also that the convergence of our approximation of topological entropy is even faster than that of Shannon's metric entropy. Shannon's metric entropy of the sequence u for the value n is defined as

$$H_{\text{met}}(u,n) = \frac{-1}{n}\sum_{w}\mu_u(w)\log(\mu_u(w))$$

where the summation is over all words of length n and $\mu_u(w)$ is the probability (frequency) of the word w appearing in the given sequence u. Thus, Shannon's metric entropy requires not just the appearance of subwords, but for the actual frequency of appearance of the subwords to converge as well. As can be seen from Definition 3, our notion of topological entropy does not require the use of the actual subword frequencies. So topological entropy will in general be more accurate than Shannon's metric entropy for shorter sequences. Accordingly, the convergence issues mentioned in Farach *et al.* (1995) (even with the clever Lempel–Ziv estimator) can be circumvented.

Furthermore, it is not difficult to show [as in Blanchard *et al.* (2000), Proposition 1.2.5] what is known as the *Variational Principle*, that is, topological entropy dominates metric entropy: for any sequence u (finite or not) and integer n

$$H_{\text{met}}(u,n) \leq H_{\text{top}}(u,n) \tag{3}$$

Thus, topological entropy retains connections to the information theoretic interpretation of metric entropy as set forth by Shannon (1948). Since topological entropy bounds metric entropy from above:

> Low topological entropy of a sequence implies that it is 'less chaotic' and is 'more structured'.

This connection to information theory is also an argument for the use of topological entropy over Rènyi continuous entropy of order α [see Vinga and Almeida (2004) for more details]. Rènyi (1961) showed that for $\alpha \neq 1$, one cannot define conditional and mutual information functions and hence Rènyi continuous entropy does not measure 'information content' in the usual sense. So while Rènyi entropy does allow for the identification of statistically significant motifs (Vinga and Almeida, 2004), one cannot conclude that higher/lower Rènyi continuous entropy for $\alpha \neq 1$ implies more/less information content or complexity in the usual sense.

Thus, LC is the only other similar measurement of sequence complexity that produces a single number representing the complexity of a sequence. Like our implementation of topological entropy, the implementation of LC contained in Troyanskaya *et al.* (2002) also runs in linear time. A comparison of our implementation of topological entropy and LC is contained in Section 4.4.

4 APPLICATION TO EXONS/INTRONS OF THE HUMAN GENOME

4.1 Method

We now apply our definition of topological entropy to the intron and exon regions of the human genome.

We retrieved the February 2009 GRCh37/ hg19 human genome assembly from the UCSC database and utilized Galaxy (Blankenberg *et al.*, 2007, 2010) to extract the nucleotide sequences corresponding to the introns and exons of each chromosome (including ChrX and ChrY). Now even though as argued above, topological entropy converges more quickly than metric entropy, one must be careful to not use this definition of topological entropy on sequences that are too short as this would lead to significant noise. For example, the UCSC database contains exons that consist of a single base and it is meaningless to attempt to measure topological entropy of such sequences. Hence, we selected the longest 100 different intron and exon sequences from each chromosome.

After ensuring that each sequence consisted only of letters from $\{A, C, T, G\}$, we then applied the approximation of topological entropy found in Definition 3 to the resulting sequences. For comparison's sake, we also applied the approximation of topological entropy to the longest 50, 200 and 400 sequences, as well as to *all* the intron and exon sequences. The salient observed features persist throughout. Though as expected, when shorter sequences are allowed, the results become noisier. These error bar plots are available in the Supplementary Material.

To investigate in more detail the relationship between regions under selective pressure and the value of topological entropy, we also selected each $5'$ and $3'$ UTR on chromosome Y that consisted of more than $4^3 + 3 - 1 = 66$ bp.

4.2 Data

Figure 7 displays the error bar plot for the longest 100 exons and introns. The error bar plots for the longest 50, 200 and 400 sequences, as well as the plot for all the intron and exon sequences, are available in the Supplementary Material.

4.3 Analysis and discussion

We first discuss the results regarding intron and exon regions. As Figure 7 demonstrates, the topological entropies of intron regions of the human genome are larger than the topological entropies of the exon regions. For example, the mean of the entropies of the introns on chromosome 21 is more than 11 SDs away from the mean of the entropy of the exons on the same chromosome. This result supports the commonly held notion that intron regions of DNA are mostly free from selective pressure and so evolve more randomly than do exon regions. We thus suggest that the observation of Karamanos *et al.* (2006), Troyanskaya *et al.* (2002), Mantegna *et al.* (1995), and Stanley *et al.* (1999) that intron entropy is *smaller* than exon entropy is due to the aforementioned finite sample effects and high-dimensionality problems related to previous implementations of entropy.

Interestingly, even though we observe that intron entropy is larger than exon entropy, the entropies of *both* regions are much lower than expected (here expectation is as calculated in Table 1). Indeed, of the longest 100 sequences, the average intron length is 180 880 and the average exon length is 2059, so according to Tables 1 and 2, we would expect the entropies to be 0.966914 and 0.933853, respectively. We find, though, that the average entropy for introns is 0.9323166 and for exons is 0.897451. Note that the largest intron sequence entropy ($H_{\text{top}} = 0.943627$ for an intron of length 1.1 Mb found on chromosome 16) is significantly lower than the expected value of 0.969921 (at least 60 SDs from the expectation). This is not too surprising considering that the expectation as calculated in Theorem 2 uses the uniform distribution. This supports the conclusion that while intron regions do evolve more randomly than exon regions, introns do not evolve uniformly randomly.

Note the disparity between the entropies of the sex chromosomes: the entropy of chromosome X in both intron and exon regions is significantly higher than in chromosome Y. In fact, the mean of chromosome X intron entropies is 3.5 SDs higher than the mean of chromosome Y intron entropies; the mean of chromosome X exon entropies is 1 SD higher than the mean of chromosome Y exon entropies. Thus, the X chromosome has intron and exon entropy similar to that of the autosomes, but chromosome Y has significantly differing exon and intron entropy. This is a particularly puzzling result considering that chromosome Y is known to have a high mutation rate and a special selection regime (Graves 2006; Wilson and Makova 2009a, b), and so one would expect the entropy of chromosome Y (both intron and exon regions) to be much higher than it is. In fact, the chromosome Y introns have the lowest mean topological entropy of any intron region across the entire genome. This would suggest that the accumulation of 'junk' DNA and the massive accumulation of retrotransposable elements mentioned in Graves (2006) have some underlying function or structure. More specifically, it appears that the intron regions in chromosome Y might fall into two categories: the truly 'junk' DNA consisting of the introns with topological entropy greater than 0.910, and the introns that have hidden structure consisting of those sequences with entropy less than 0.910. We present in Figure 4 a histogram of the topological entropy on chromosome Y demonstrating the distinction between the two categories.

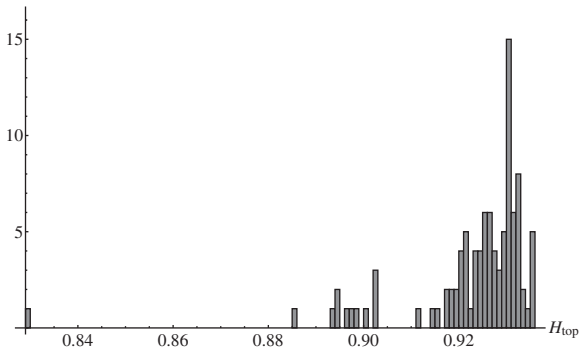

Fig. 4. Histogram of topological entropy of introns in chromosome Y.

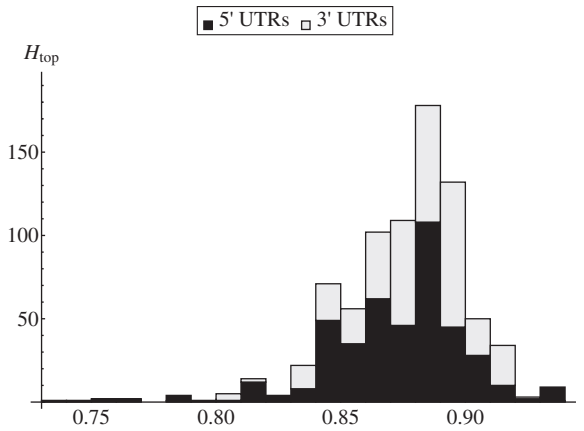

Fig. 5. Histogram of topological entropy for 5′ and 3′ UTRs in chromosome Y.

Remaining on chromosome Y, we now present evidence that topological entropy can be used to detect sequences that are under selective pressure. Note that Siepel *et al.* (2005) showed that both 5′ and 3′ UTRs are among the most conserved elements in vertebrate genomes. Thus, one would expect that the topological entropy of these regions would be very low (as this is indicative of a high degree of structure). As indicated in Figure 6, the entropy of both the 5′ and 3′ region are low in comparison to the entropy of the intron and exon regions across the autosomes. Compare, for example, Figure 4 and 5. In fact, the mean of the topological entropy of the 5′ and 3′ UTRs (0.871545 ± 0.0290619 and 0.879163 ± 0.0219371) are lower than the mean entropy of *any* intron or exon region across every chromosome. The lowest mean topological entropy for an autosome is 0.927802 ± 0.00539 on chromosome 19, this is more than 9 SDs *higher* than the mean of topological entropy for either the 3′ or 5′ UTRs. This lends support to the assertion that topological entropy can be used to detect functional regions and regions under selective constraint.

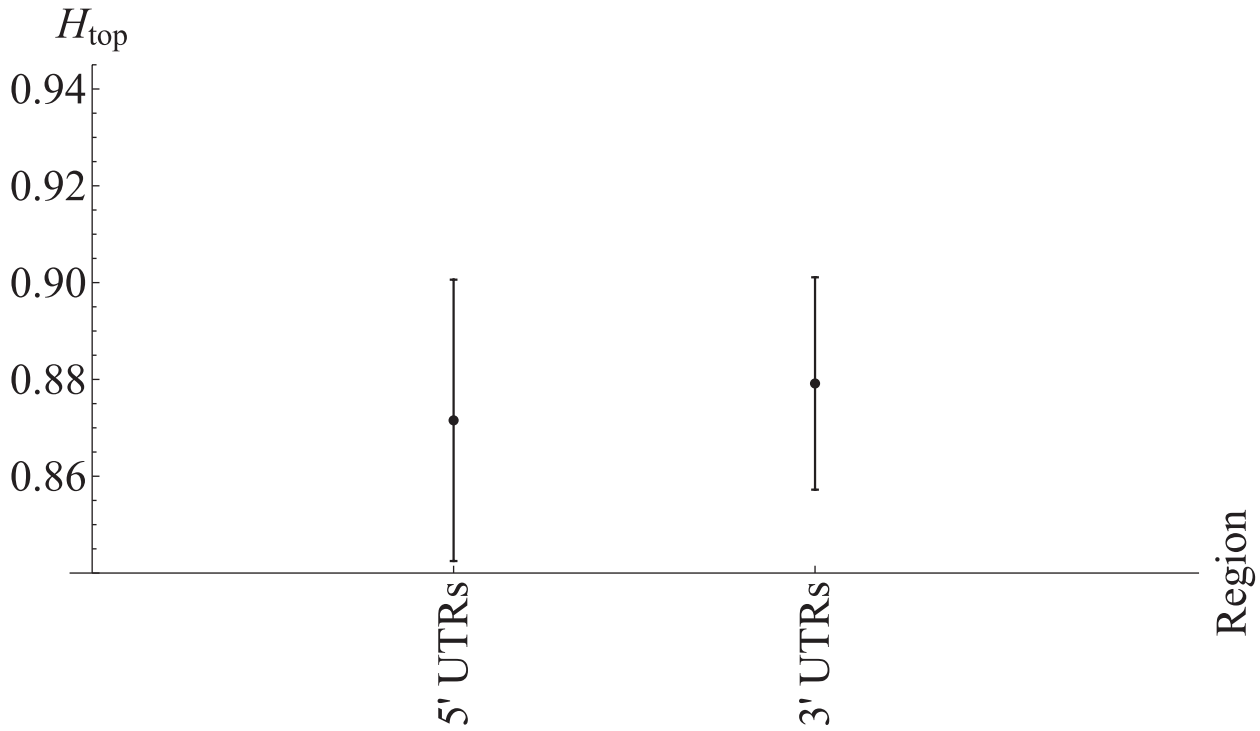

Fig. 6. Error bar plot of chromosome Y 5′ and 3′ UTRs longer than 66 bp long.

4.4 Comparison to LC

As mentioned in Section 3.1, LC is the only other similar measurement of sequence complexity that produces a single number to represent the complexity of a sequence. We applied the algorithm described in Troyanskaya *et al.* (2002) and written by Larsson (1999) to the same dataset contained in Section 4.1 of this article. To obtain directly comparable results, we used a window size as big as the given sequence is long. As can be seen in the Supplementary Material, LC does distinguish between introns and exons to an extent, though not to the same quality of resolution as that of topological entropy (compare with Fig. 7). For example, while topological entropy consistently measures introns as more random than exons, LC does not. This discrepancy is most likely due to linguistic complexity being effectively utilized (Troyanskaya *et al.*, 2002) as a sliding window method to detect repetitive motifs, not as a holistic measure of sequence information content. So we also applied LC using a sliding window of 2000 bp, taking the average value of LC on a given sequence, and then averaging on a given chromosome. Using the sliding window, LC does give a higher value to introns than to exons (except on chromosome 5). While the separation between the LC of introns and exons becomes more pronounced, the resolution is still not nearly as clear as with topological entropy since a large amount of error persists. The LC values among introns and exons are well within 1 SD of each other across the entire genome.

5 CONCLUSION

This implementation of topological entropy is free from issues that other implementations have encountered. Namely, this definition allows for the comparison of sequences of different length and does not suffer from multidimensionality complications. Since this definition supplies a single value to characterize the complexity of a sequence, it is much more capable of being mathematically analyzed. Beyond measuring the complexity or simplicity of a sequence, we presented evidence that our approximation to topological entropy might detect functional regions and sequences free from or under selective constraint. The speed and simplicity of this implementation of topological entropy makes it very suitable for utilization in detecting regions of high/low complexity. For example, we observe the novel phenomena that the introns on chromosome Y have atypically low and bimodal entropy, possibly corresponding to random sequences and sequences that posses hidden structure or function.

ACKNOWLEDGEMENTS

The author would like to thank Manfred Denker, Kateryna Makova and Francesca Chiaromonte for their assistance and fruitful discussion regarding this article.

Funding: National Science Foundation (grant number DMS-1008538).

Conflict of Interest: none declared.

REFERENCES

Blanchard,F. *et al.* (eds) (2000) *Topics in Symbolic Dynamics and Applications*. London Mathematical Society Lecture Note Series 279. Cambridge University Press, Cambridge, UK.

Blankenberg,D. *et al.* (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.*, **17**, 960–964.

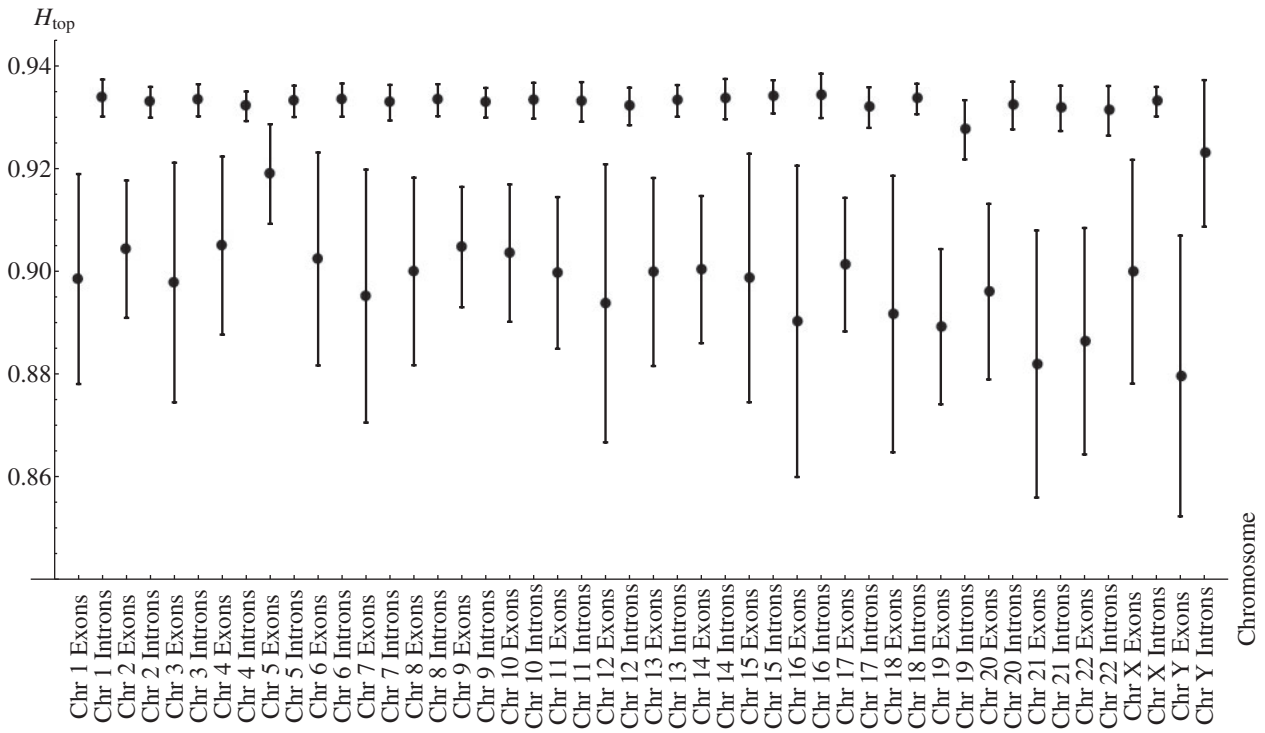

Fig. 7. Error bar plot of average topological entropy for the longest 100 introns and exons in each chromosome.

Blankenberg,D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protocol. Mol. Biol.*, **19**, 1–21

Colosimo,A. and de Luca,A. (2000) Special factors in biological strings. *J. Theor. Biol.*, **204** , 29–46.

Crochemore,M. and Vèrin,R. (1999) Zones of low entropy in genomic sequences. *Comput. Chem.*, **23**, 275–282.

Farach,M. *et al.* (1995) On the entropy of DNA: algorithms and measurements based on memory and rapid convergence. In *Proceedings of the sixth annual ACM-SIAM symposium on discrete algorithms.* SIAM, Philadelphia, PA, pp. 48–57.

Gabrielian,A. and Bolshoy,A. (1999) Sequence complexity and DNA curvature. *Computers & Chemistry*, **23**, 263–274.

Gheorghiciuc,I. and Ward,M.D. (2007) On correlation polynomials and subword complexity. In *Conference on Analysis of Algorithms*, *Discrete Mathematics and Theoretical Computer Science Proceedings AH*. Nancy, France, pp. 1–18.

Graves,J.A.M. (2006) Sex chromosome specialization and degeneration in mammals. *Cell*, **124**, 901–914.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.) *Mammalian Protein Metabolism.* Academic Press, New York, pp. 21–132.

Karamanos,K *et al.* (2006) Statistical compressibility analysis of DNA sequences by generalized entropy-like quantities: towards algorithmic laws for Biology? *Proc. 6th WSEAS Int. Conf. Appl. Informat. Commun.*, **18**, 481–491.

Kirillova,O.V. (2000) Entropy concepts and DNA investigations. *Phys. Lett. A*, **274**, 247–253.

Larsson,N.J. (1999) Structures of String Matching and Data Compression. PhD Thesis, Lund University, Sweden.

Mantegna,R.N. *et al.* (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E*, **52**, 2939–2950.

Ornstein,D. and Weiss,B. (2007) Entropy is the only finitely observable invariant. *J. Mod. Dyn.*, **1**, 93–107.

Rènyi,A. (1961) On measures of information and entropy. In *Proceedings of the 4th Berkely Symposium on Mathematical Statistics and Probability*, Vol. I. University of California Press, Berkely, CA, pp. 547–561.

Schmitt,A.O. and Herzel,H. (1997) Estimating the entropy of DNA sequences. *J. Theor. Biol.*, **1888**, 369–377.

Shannon,C.E. (1948) A Mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Stanley,H.E. (1999) Scaling features of noncoding DNA. *Phys. A*, **273**, 1–18.

Troyanskaya,O.G. *et al.* (2002) Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, **18**, 679–688.

Vinga,S. and Almeida,J.S. (2004) Rènyi continuous entropy of DNA sequences. *J. Theor. Biol.*, **231**, 377–388.

Wilson,M.A. and Makova,K.D. (2009a) Genomic analyses of sex chromosome evolution. *Annu. Rev. Genome Hum. Genet.*, **10**, 333–354.

Wilson,M.A. and Makova,K.D. (2009b) Evolution and survival on eutherian sex chromosomes. *PLoS*, **5**, e1000568.