A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences

Shuilin Jin[1]☯, Renjie Tan[2]☯, Qinghua Jiang[3], Li Xu[4], Jiajie Peng[2], Yong Wang[1]*, Yadong Wang[2]*

[1] Department of Mathematics, Harbin Institute of Technology, Harbin, Heilongjiang, China, [2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China, [3] School of Life Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China, [4] College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China

Abstract

Topological entropy is one of the most difficult entropies to be used to analyze the DNA sequences, due to the finite sample and high-dimensionality problems. In order to overcome these problems, a generalized topological entropy is introduced. The relationship between the topological entropy and the generalized topological entropy is compared, which shows the topological entropy is a special case of the generalized entropy. As an application the generalized topological entropy in introns, exons and promoter regions was computed, respectively. The results indicate that the entropy of introns is higher than that of exons, and the entropy of the exons is higher than that of the promoter regions for each chromosome, which suggest that DNA sequence of the promoter regions is more regular than the exons and introns.

Citation: Jin S, Tan R, Jiang Q, Xu L, Peng J, et al. (2014) A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences. PLoS ONE 9(2): e88519. doi:10.1371/journal.pone.0088519

Editor: Zhongxue Chen, Indiana University Bloomington, United States of America

Received October 11, 2013; **Accepted** January 6, 2014; **Published** February 12, 2014

Funding: This work is supported by the China Natural Science Foundation (Grant No. 11301110, No. 61201084, No. 61102149 and No. 61173085), China Postdoctoral Science Foundation (Grant No. 2013M541346), the State Scholarship Fund from the China Scholarship Council (CSC) (Grant No. 201203070290 and No. 201206125024), the Program for Innovation Research of Science in Harbin Institute of Technology and the Fundamental Research Funds for the Central Universities (Grant No. HEUCF100604). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ydwang@hit.edu.cn (Yadong Wang); mathwy@hit.edu.cn (Yong Wang)

☯ These authors contributed equally to this work.

Introduction

The first concept of entropy was introduced by Shannon[8] as a measure of the complexity of a set of symbols, which can be formulated in mathematical form as:

$$H=-\sum_{i=1}^{n}p_i\log(p_i)$$

where p_i is the probability of the i-th symbol. Since then the notions of entropy appeared in many forms, such as metric entropy, topological entropy, Kolmogorov-Sinai entropy and Rènyi [7] entropy. All of the concepts were focused on one purpose: the "quantitative" description of the complexity or simplicity of a set of symbol dynamics.

The complexity of DNA sequences, as a special kind of symbol dynamics which is composed of A,C,G,T, can be measured by the entropy. Kirillova [5] computed DNA sequences of different organisms by the topological and metric entropies. Vinga and Almeida [9] introduced Rènyi's quadratic entropy to evaluate the randomness of DNA sequences. Zhao F, Yang H and Wang B [10] investigated the complexity of human promoter sequences by a diffusion entropy. Bose and Chouhan [3] studied the super-information of the DNA sequence. Recently, Koslicki [6] introduced a topological entropy for finite sequences and showed the complexity of introns is higher than that of exons for each chromosome.

In this paper, a generalized topological entropy is introduced. At the same time, the relationship between the topological entropy and the generalized topological entropy is compared, which shows the topological entropy is a special case of the generalized entropy. The use of generalized topological entropy removes high-dimensional problems. This definition can get the complexity of sequences of different length. At last, we apply the generalized topological entropy to human genome to compute the complexity of introns, exons and promoter regions.

Methods

Let ω be a sequence of DNA with length $|\omega|$, $p_\omega(n)$ be the number of different n-length subwords that appear in ω. If the sequence is infinite, then the topological entropy is defined as:

Definition 1. For an infinite sequence ω formed over A,C,G,T, the topological entropy is

$$H_T(\omega)=\lim_{n\to\infty}\frac{\log_4(p_\omega(n))}{n}.$$

Take a symbol sequence $\omega=CGCGCGCG\cdots$ as an example. It is easily seen, for any n, the different sequence with length n is 2, so the topological entropy of the DNA sequence $CGCGCGCG\cdots$ is:

$$H_T(\omega) = \lim_{n\to\infty} \frac{2}{n} = 0.$$

However, the length of DNA sequence is finite, by Definition 1, the complexity is zero as n tends to infinity. Colosimo and Luca [4] showed the precise description of the shape of the complexity function, and then Koslicki defined an approximation of topological entropy $H_K(\omega) = \frac{\log_4(p_\omega(n))}{n}$.

The spirit of Koslicki's method is to find a unique n, which satisfies $4^n + n - 1 \le |\omega| < 4^{n+1} + (n+1) - 1$, and then the complexity is decided by the subwords of length n. However, by comparing the topological entropies of DNA sequences, we find the complexities of subwords, which are shorter than n, are also important. The following Figure 1 is the topological entropy of the promoter regions (2000 bp upstream before the transcription start site) before gene WASH7P and TMCO4 on chromosome 1.

Following Koslicki's definition, the complexity of the promoter region before TMCO4 is higher than WASH7P. However, considering the complexity of subwords, the fact is not the case. As a matter of fact, **the complexities of subwords all contribute to the complexity of the sequence**. Based on the idea of overall consideration, we give the following definition.

Definition 2. Let ω be an infinite sequence formed over A,C,G,T, the generalized topological entropy is

$$H_{\mathrm{GT}}(\omega) = \lim_{k\to\infty} \sum_{i=1}^{k} \alpha_i^{(k)} \frac{\log_4(p_\omega(i))}{i}$$

where

$$\alpha_i^{(k)} = \begin{cases} a_i & i \le k \\ 0 & i > k \end{cases} \quad \text{and} \quad \sum_{i=1}^{k} a_i = 1, \quad a_i \ge 0,$$

satisfies for any $\varepsilon > 0$, for any I, there exists K, such that for all $k \ge K$, $\sum_{i=I+1}^{\infty} \alpha_i^{(k)} > 1 - \varepsilon$.

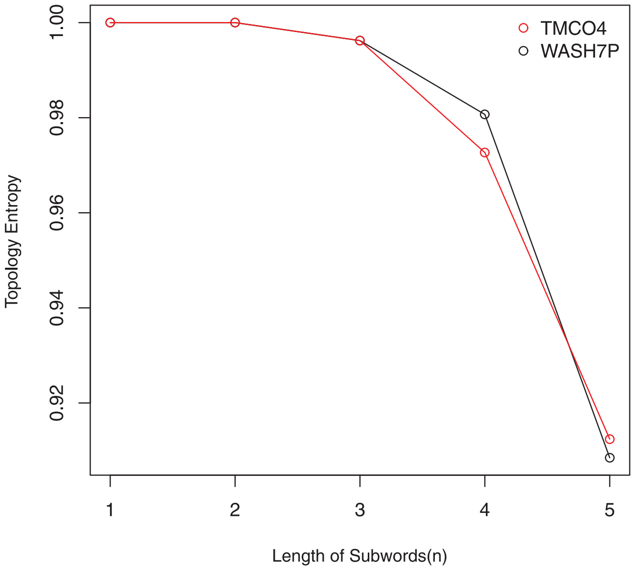

Figure 1. The topological entropy of the promoter regions before gene WASH7P and TMCO4. Notice that the unique number which satisfies $4^n + n - 1 \le 2000 < 4^{n+1} + (n+1) - 1$ is 5. The topological entropy of promoter regions before gene WASH7P and TMCO4 are 0.908480839 and 0.912412131.
doi:10.1371/journal.pone.0088519.g001

For infinite sequences formed over A,C,G,T, then $H_{GT}(\omega) = H_T(\omega)$. (Appendix S1)

By considering the complexity of finite sequence, we give the following generalized topological entropy.

Definition 3. Let ω be a finite sequence of length $|\omega|$. Let n_ω be the unique integer such that $4^n + n - 1 \le |\omega| < 4^{n+1} + (n+1) - 1$.

Then for $\omega_1^{4^n+n-1}$ the first $4^n + n - 1$ letters of ω and $k \le n_\omega$,

$$H_{n_\omega}^{(k)}(\omega) = \frac{1}{k} \sum_{i=n_\omega-k+1}^{n_\omega} \frac{\log_4(p_\omega(i))}{i}$$

Notice that the generalized topological entropy by Definition 3 satisfies the following four important properties mentioned by Koslicki [6].

(1) $0 \le H_{n_\omega}^{(k)}(\omega) \le 1$;

(2) $H_{n_\omega}^{(k)}(\omega) \approx 0$ if and only if ω is highly repetitive;

(3) $H_{n_\omega}^{(k)}(\omega) \approx 1$ if and only if ω is highly complex;

(4) For different length sequences ω,v and $k \le min\{n_\omega, n_v\}$, $H_{n_\omega}^{(k)}(\omega)$and $H_{n_v}^{(k)}(v)$ can be comparable.

It is easily seen for infinite sequence, $H_{n_\omega}^{(k)}(\omega)$ is a special case of $H_{\mathrm{GT}}(\omega)$. Moreover, the topological entropy defined by Koslicki satisfies $H_K(\omega) = H_{n_\omega}^{(1)}(\omega)$.

Applications to Human Genome

Data

We retrieved the hg19 human genome assembly from the UCSC database and utilized Galaxy(Blankenberg [1–2]) to extract the nucleotide sequences of the introns, exons and promoter regions of each chromosome. The sequences that are too short would lead to significant noise. For example, the UCSC database contains exons that are only one base pair long and it is trivial to measure the complexity of such sequences. We selected randomly 100 different promoter sequences (2000 bp upstream before the transcription start site) from each chromosome, and repeated this procedure for 100 times and computed the average complexity of promoter sequences.

Results and discussions

We used the generalized topological entropy to compute the complexity of introns, exons and promoter regions of the human genome by non-overlapping windows algorithm by Koslicki.

Figure 2 displays the complexity of human introns, exons and promoter regions by $H_{n_\omega}^{(3)}(\omega)$. Here we took the mean of the generalized topological entropies of 100 different introns and exons sequences from each chromosome. We did ANOVA of the complexity differences among introns, exons and promoters on each chromosome and found the differences are statistical significant. (p-value<0.005, Appendix S2)

It is seen by Figure 2, the generalized topological entropy of introns is higher than exons for each chromosome, which demonstrates the structure of introns is more complex than exons. This is reasonable

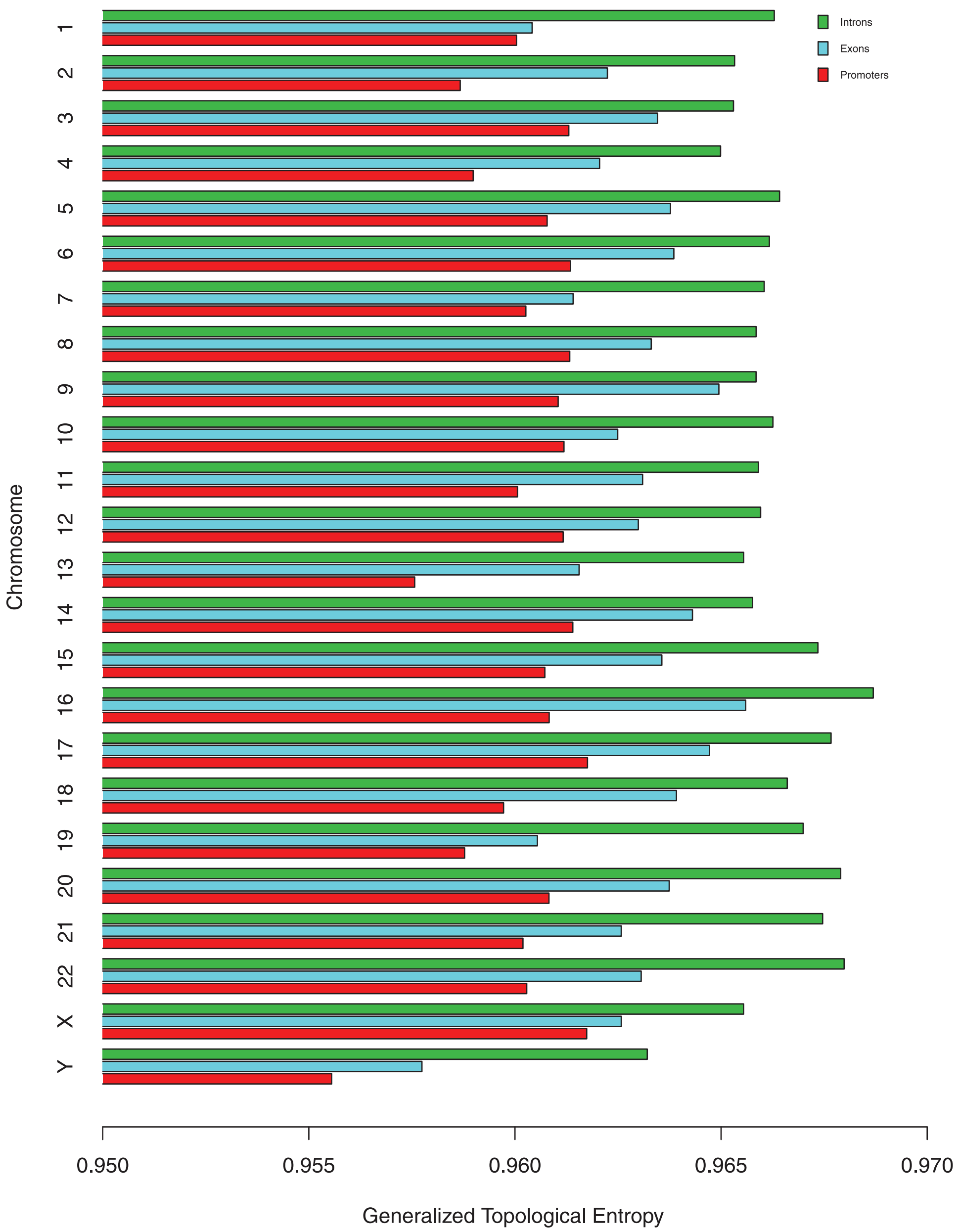

Figure 2. Generalized topological entropies of introns, exons and promoter regions.
doi:10.1371/journal.pone.0088519.g002

due to the fact the introns of DNA sequences are free from selective pressure and so evolve more randomly than exons.

Note that promoter regions are among the most conserved elements in Eukaryotic genomics, which consist of the TATA box, CAAT box, GC-enriched region and so on. Thus one would expect that the generalized topological entropy of these regions would be very low. As shown in figure 2, the mean of the generalized topological entropies of 100 different promoter regions is lower in comparison to the mean of the generalized topological entropies of 100 different introns and exons for each chromosome, which suggests the generalized topological entropy can be used to detect functional regions and regions under selective constraint.

Conclusions

The generalized topological entropy has two advanced features. Theoretically, the definition of generalized topological entropy is a complete form of topological entropy. Practically, the use of the generalized topological entropy allows comprehensive analysis of the complexity of DNA sequences, which counts for almost all the complexities of the subwords. Besides measuring the complexity or simplicity of sequences, the generalized topological entropy can be used to detect functional regions and regions under selective constraint.

Supporting Information

Appendix S1 Proof of the generalized topological entropy.
(DOC)

Appendix S2 Complexity differences among introns, exons and promoters.
(DOC)

References

1. Blankenberg D, Taylor J, Schenk I, He J, Zhang Y, et al. (2007) A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. Genome Res 6: 960–964.
2. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Current Protocols in Molec Biol 19: 1–21.
3. Bose R, Chouhan S (2011) Alternate measure of information useful for DNA sequences. Phys Rev E 83: 051918.
4. Colosimo A, de Luca A (2000) Special factors in biological strings. J Theor Biol 204: 29–46.
5. Kirillova OV (2000) Entropy concepts and DNA investigations. Phys Lett A 274: 247–253.
6. Koslicki D (2011) Topological entropy of DNA sequences. Bioinformatics 27: 1061–1067.
7. Rènyi A (1961) On measures of information and entropy: In Proceedings of the 4th Berkely Symposium on Mathematical Statistics and Probability, Berkely: University of California Press.
8. Shannon CE (1948) A Mathematical theory of communication. Bell Syst Tech J 27: 379–423.
9. Vinga S, Almeida JS (2004) Rènyi continuous entropy of DNA sequences. J Theor Biol 231: 377–388.
10. Zhao F, Yang H, Wang B (2007) Complexity of human promoter sequences. J Theor Biol 247: 645–649.

Author Contributions

Conceived and designed the experiments: SJ Yadong Wang Yong Wang. Performed the experiments: SJ RT. Analyzed the data: QJ LX. Contributed reagents/materials/analysis tools: JP Yadong Wang. Wrote the paper: SJ Yong Wang Yadong Wang.