Comments on "Informational laws of genome structures"

Martín Andrade-Restrepo (martin.andrade@urosario.edu.co)
University of el Rosario
Carlos É. Alvarez
University of el Rosario

Research Article

Keywords:

Posted Date: March 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1416535/v1

Comments on "Informational laws of genome structures"

Martín Andrade-Restrepo[1,*,+] **and Carlos E. Álvarez**[1,+]

[1]School of Engineering, Science and Technology, University of el Rosario, Bogotá, Colombia
[*]martin.andrade@urosario.edu.co
[+]these authors contributed equally to this work

ABSTRACT

In this paper, we expose possible flaws in some of the proofs of the mathematical results presented by *Bonnici and Manca, 2016* and suggest some corrected proofs of two of their results. We show that the possibility of overlap between k-mers given a fixed value of k (minimum length such that all k-mers are hapaxes) was not taken into account in several steps of the proof in Lemma 2, which makes the theorem false. For this we show a counterexample to the Lemma. In addition, we correct the proofs of Lemma 1 and Proposition 3. In the former case, we explain why the argument presented is incomplete and provide a correct proof. In the case of the latter, we show with a counter example that their claim is incorrect without explicitly mentioning the correct ranges of k over which it holds. We also provide a more general proof which accounts for this correction. We remark that although some of these comments invalidate the mathematical basis of their results, there is a possibility that the associated error becomes small in the limit $n \gg k$, hence allowing for the apparent agreement with the data and the simulations.

Introduction

A k-mer is a substring α of length k from a genome $\mathbb{G}$ of length $n = |\mathbb{G}| \geq k$, and the frequency distribution of k-mers in genomes is frequently used to characterize them[1–4]. A particularly useful parameter of the distribution is its entropy, defined as

$$E_k(\mathbb{G}) = -\sum_{\alpha \in D_k(\mathbb{G})} p(\alpha) \lg_2 p(\alpha), \tag{1}$$

where E_k is the entropy of the distribution of k-mers, $p(\alpha)$ is the probability of finding k-mer α in $\mathbb{G}$ and $D_k(\mathbb{G})$ is the set of all k-mers present in $\mathbb{G}$. The work from *Bonnici and Manca*[5] proposes that there is a preferential value of k for which particular indexes can be derived, which follow a set of "informational laws".

The first result obtained is that, when $k = mrl(\mathbb{G}) + 1$, where $mrl(\mathbb{G})$ is the length of the longest repeats of $\mathbb{G}$, the following relation

$$LG(\mathbb{G}) < E_{2LG}(\mathbb{G}) < 2LG(\mathbb{G}), \tag{2}$$

where $LG(\mathbb{G}) = \lg_4(n)$, was observed to hold for a set of 70 genomes studied. The upper bound in eq. (2) is then proven in Proposition 3, to which we address some comments at a later section, while no proof is presented for the lower bound.

From (2) the definitions for *entropic component* value

$$EC(\mathbb{G}) = E_{2LG}(\mathbb{G}) - LG(\mathbb{G}), \tag{3}$$

anti-entropic component value

$$AC(\mathbb{G}) = 2LG(\mathbb{G}) - E_{2LG}(\mathbb{G}), \tag{4}$$

and *lexical index*

$$LX(\mathbb{G}) = \frac{|\mathbb{G}|}{|D_m(\mathbb{G})|}, \tag{5}$$

where $m = 2LG$, are presented.

The authors state that the identification of the parameters, defined previously, relies on two lemmas and one proposition: **Lemma 1**, **Lemma 2** and **Proposition 3**. Hence, we consider essential to share what we believe are some inaccuracies and faults made in their proofs. We outline our comments in the following sections

Results

Corrections to Lemma 1

Lemma 1 from the states: "*Given a genome $\mathbb{G}$ of length n, if $k = mrl(\mathbb{G})+1$, then $E_k(\mathbb{G})$ is the maximum value that E_k can reach in the class of all possible genomes of length n*".

Here, "*...$mrl(\mathbb{G})$ is the length of the longest repeats of $\mathbb{G}$ and $mrl(\mathbb{G})+1$ is the minimum length, such that k-mers with k greater than $mrl(\mathbb{G})$ are all hapaxes.*"

The proof presented in the article uses the Equipartition Principle. This claim is incorrect for the following reason. By reducing the value of k from $mrl(\mathbb{G})+1$ to a value $k < mrl(\mathbb{G})+1$, there exists at least one k-mer which is no longer a hapax (it occurs more than once in the genome). Hence, the distribution of the k-mers that appear in $\mathbb{G}$ for each of the available slots is no longer necessarily uniform, which could (mistakenly) be used to justify a decrease in the entropy based on the Equipartition Principle, if one would not consider the changes in the size of the set of possible outcomes. However, the size of the set of k-mers that appear in $\mathbb{G}$ can increase as k drops. For instance, consider the sequence *AGCTAGCT*. Here, the entropy for $k = mrl(\mathbb{G})+1 = 5$ is calculated over 4 possible outcomes (the four 5-mers). However, for $k = 2$ the entropy is calculated over seven 2-mers. If the size of the set of possible outcomes of an experiment is not kept constant, the entropy can increase by diverging from uniform distributions while increasing the size of such set. As an example, take the uniform distribution over a set of two possible outcomes (the probabilities for both are $1/2$). It is easy to verify that the entropy is $\lg_2 2 = 1$. If the set of possible outcomes has size 3, and the probabilities are $\{1/6, 2/6, 3/6\}$ (the distribution is not uniform), the entropy is approximately 1.46. Hence, arguing that the entropy is maximal for $k = mrl(\mathbb{G})+1$, given that the distribution is uniform, is not enough to prove the lemma.

We provide a corrected version of the proof below.

Proof. Let $n \geq 1$ be an integer and $\mathbb{G}$ a random genome of length n. Let $k = mrl(\mathbb{G})+1$, $n \geq k \geq 1$, be the minimum length such that all k-mers occuring in $\mathbb{G}$ are hapaxes and let $b_2 := n-k+1$; the number of possible slots in the genome where each k-mer can occur. It is easily verified that:

$$E_k(\mathbb{G}) = -\sum_{\alpha \in D_k(\mathbb{G})} p(\alpha) \lg_2 p(\alpha) = b_2 \left(\frac{1}{b_2}\right) \lg_2(b_2) = \lg_2(b_2).$$

First consider the case where s is an integer such than $k \leq s \leq n$. Since all s-mers occuring in $\mathbb{G}$ are hapaxes, their sampling distribution at each position of $\mathbb{G}$ is uniform. The number of s-mers occuring in $\mathbb{G}$ is $b_1 := n-s+1 \leq b_2$. Hence,

$$E_s(\mathbb{G}) = -\sum_{\alpha \in D_k(\mathbb{G})} p(\alpha) \lg_2 p(\alpha) = b_1 \left(\frac{1}{b_1}\right) \lg_2(b_1) = \lg_2(b_1) \leq \lg_2(b_2) = E_k(\mathbb{G}).$$

Now, suppose $k > 1$ and let s be an integer such that $1 \leq s < k$. Since $mrl(\mathbb{G}) = k-1 > 0$, there is at least one $(k-1)$-mer that is a repeat in $\mathbb{G}$. Moreover, there are at least $(k-s)$ s-mers which are also repeats (the number of s-mers which occur in the $(k-1)$-mer which is a repeat in $\mathbb{G}$). Assuming that no other s-mers besides those inside We will consider two cases.

Case 1: Suppose the two $(k-1)$-mers which are repeats do not overlap (Fig. 1 panel a).

It is easy to see that the entropy $E_s(\mathbb{G})$ is maximal if, besides the s-mers contained inside the two repeats of length $k-1$, no other s-mer occurs more than once in $\mathbb{G}$ (since this way the size of the dictionary is maximal) and if each of the repeats of length s occurs only twice (by the equipartition principle since this way one is closest from a uniform distribution). In other words, the entropy is maximal if the $(n-s+1-2(k-s))$ s-mers, not contained in the longest repeats appear only once and the $(k-s)$ contained in one of the longest repeats appear only twice (once in each of the longest repeats). Hence,

$$\begin{aligned} E_s(\mathbb{G}) &\leq -(n-s+1-2(k-s)) \left(\frac{1}{n-s+1}\right) \lg_2 \left(\frac{1}{n-s+1}\right) - (k-s) \left(\frac{2}{n-s+1}\right) \lg_2 \left(\frac{2}{n-s+1}\right) \\ &= (n-s+1-2(k-s)) \left(\frac{1}{n-s+1}\right) \lg_2 (n-s+1) + \left(\frac{2(k-s)}{n-s+1}\right) (\lg_2(n-s+1) - \lg_2(2)) \\ &= \frac{n-s+1}{n-s+1} (\lg_2(n-s+1)) - 2\frac{k-s}{n-s+1} \lg_2(2) \end{aligned}$$

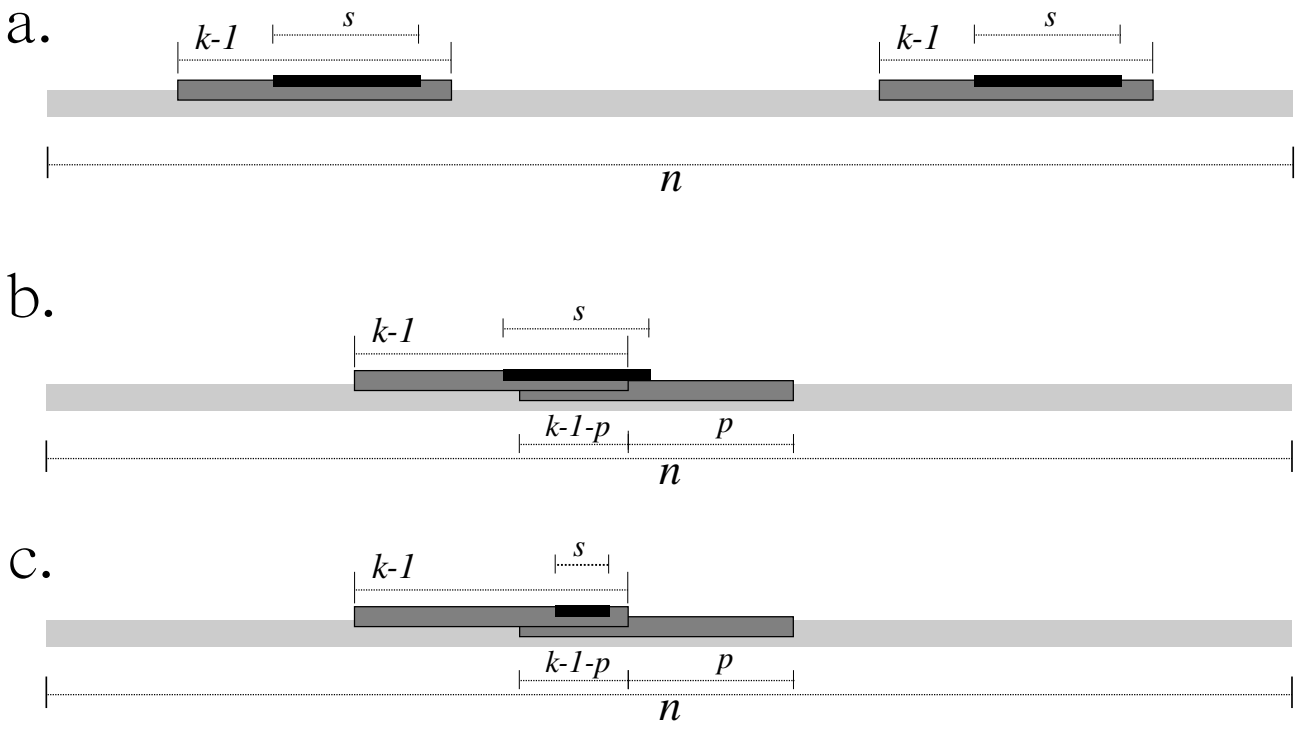

Figure 1. Different cases considered in the second part of the proof of Lemma 1. Case 1 (panel a): the repeated $(k-1)$-mers do not overlap. Case 2.1 (Panel b): there is an overlap between the repeated $(k-1)$-mers, and $s > k-1-p$. Case 2.2 (Panel c): there is an overlap between the repeated $(k-1)$-mers, and $s \leq k-1-p$.

$$= (\lg_2(n-s+1)) - 2\frac{k-s}{n-s+1} = \lg_2(b_1) - 2\frac{b_1-b_2}{b_1},$$

where, once more, $b_1 = n-s+1$, $b_2 = n-k+1$.

W must then prove that:

$$\lg_2(b_1) - 2\frac{b_1-b_2}{b_1} \leq \lg_2(b_2),$$

or, equivalently,

$$\lg_2\left(\frac{b_1}{b_2}\right) \leq 2\frac{b_1-b_2}{b_1}.$$

taking $x = b_1/b_2 > 1$, the last inequality (yet to be proven) becomes:

$$\lg_2(x) \leq 2\left(1-\frac{1}{x}\right).$$

This inequality is satisfied if $1 \leq x \leq 2$. Since, $s < k$, $x > 1$. Moreover, since there is no overlap of the repeats of length $(k-1)$,

$$2(k-1) \leq n,$$

which occurs if and only if

$$2k-n-2-s \leq -s \leq -1.$$

This last inequality implies that

$$2k-n-s-1 = n-s+1-2(n-k+1) \leq 0,$$

or, equivalently

$$b_1 \leq 2b_2,$$

which implies that $x \leq 2$.

Case 2: Suppose there is an overlap between the two $(k-1)$-mers which are repeats (Fig. 1 panels b and c). Let $p > 1$ be the length of the region in each of the repeated $(k-1)$-mers left outside of the overlap (which will have length $(k-1)-p$). We consider two subcases.

Case 2.1: Suppose first that $s > (k-1)-p$ (s larger than the size of the overlap). This case is presented in Fig. 1 panel b.

There are $((k-1+p)-s+1)$ s-mers cointained inside the two (longest) repeats of size $k-1$. All of this s-mers will appear at least twice, except those that are not fully contained inside either one of the longest repeats (see Fig. 1 panel b for an example of such an s-mer). The number of such s-mers is $(s-((k-1)-p)-1) \geq 0$, since this is the number of ways in which you can distribute the surplus of an s-mer covering the overlap in such a way that it surpasses it in both directions. Hence, we can say that at least

$$(k-1+p)-s+1-(s-((k-1)-p)-1) = 2(k-s)$$

s-mers will appear two or more times. Once again, the entropy is maximized if this s-mers appear only twice and if they are the only ones that appear more than once. Hence, once more

$$E_s(\mathbb{G}) \leq -(n-s+1-2(k-s))\left(\frac{1}{n-s+1}\right)\lg_2\left(\frac{1}{n-s+1}\right) - (k-s)\left(\frac{2}{n-s+1}\right)\lg_2\left(\frac{2}{n-s+1}\right)$$

$$= (n-s+1-2(k-s))\left(\frac{1}{n-s+1}\right)\lg_2(n-s+1) + \left(\frac{2(k-s)}{n-s+1}\right)(\lg_2(n-s+1)-\lg_2(2))$$

$$= (\lg_2(n-s+1)) - 2\frac{k-s}{n-s+1} = \lg_2(b_1) - 2\frac{b_1-b_2}{b_1}.$$

Once more,

$$\lg_2(b_1) - 2\frac{b_1-b_2}{b_1} \leq\leq \lg_2(b_2)$$

if and only if $b_2 \leq b_1 \leq 2b_2$. Again, since $s < k$, $b_2 \leq b_1$. Also, since $s > k-1-p$,

$$b_1 = n-s+1 < n-(k-1-p)+1 = n-k+1+(p+1) = b_2+(p+1).$$

Or

$$b_1 \leq b_2 + p.$$

Now, $p = [(k-1)+p]-k+1$; the number of k-mers inside the region covered by the two longest repeats. Hence, $p \leq b_2$, the number of k-mers in the entire genome and $b_1 \leq 2b_2$.

Case 2.2: Suppose now that $s \leq (k-1)-p$ (s smaller or equal than the size of the overlap). This case is presented in Fig. 1 panel c. We have that every s-mer inside the region covered by the two longest repeats will appear at least twice since every s-mer is fully contained in one of the repeats. Moreover, every s-mer in this region will be either one of the first p s-mers in it (from left to right) or a copy of one of the first p s-mers. This is due to the fact that this set of s-mers will appear again starting from position $p+1$ and so on. Hence, there are at most $[n-s+1-((k-1+p)-s+1)]+p = n-k+1$ different s-mers appearing in $\mathbb{G}$. In other words, the maximum number of different s-mers will be the number of k-mers appearing in $\mathbb{G}$. However, while the distribution of the k-mers is uniform (each one appears once), that of the s-mers will not be uniform (since there are some of them which occur more than once). Hence,

$$E_s(\mathbb{G}) \leq E_k(\mathbb{G}).$$

□

Counter example to Lemma 2

Lemma 2 reads: *If R is a random genome of length n, then*

$$\lfloor \lg_2(n) \rfloor - 1 \leq mrl(R) + 1 \leq \lceil \lg_2(n) \rceil. \tag{6}$$

The proof of this lemma presented in the article states that a k-mer has a probability $(n-k+1)/4^k$ to occur in a genome of length n, generated randomly from a uniform distribution. This is not correct, no matter how it is interpreted, and can be disproved by taking into account the presence of an overlap between the occurring k-mers.

If this probability is calculated after fixing $k = mrl(R)+1$, a simple counter example can be constructed as shown next. Assume we have $n = 3$ and $mrl(R) = 1$. That is, $k = 2$. Then the random genome R can only be one of those which have two equal monomers and a 3rd one which is different (for example ACA). The other monomer has to be different because $k = 2$. Hence, we are left with 36 possible genomes instead of $4^3 = 64$ possibilities. Moreover, the probability of encountering any (possible) 2-mer, say AC, in such a genome will be $4/36 = 1/9$ instead of $2/64 = 1/32$.

Furthermore, the probability that any 2-mer, say AC appears in a random genome of length 3 is also not equal to $2/64 = 1/32$. If we define the events $A_1 = \{\text{AC appears in the first possible position}\}$ and $A_2 = \{\text{AC appears in the second possible position}\}$, then the two events are disjoint and the probability of AC occurring in the random genome:

$$P(A_1) + P(A_2) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8} \neq 1/32.$$

Finally, the probability that AC appears in a genome of length 3 only once, given that it is appears in such genome, is equal to 1 which is also different from $1/32$. Hence, no matter how the statement presented in the proof is interpreted, it is always invalid.

This result invalidates the proof in the article.

It is also easy to find a counter example to the right hand side inequality in (6). Assume $R = \text{TCTTCTTCAG}$, so that $n = 10$. In this case the largest repeat is TCTTC and $mrl(R)+1 = 6$, which is greater than $\lceil \lg_2(n) \rceil = 4$. It might be the case that for large values of n the inequality holds on average, which would have to be proven, but it certainly does not hold in general.

Correction to Proposition 3

Proposition 3 reads: *In the class of genomes of length n, for every $k < n$, the following relation holds*

$$E_k(\mathbb{G}) < \lg_2(n). \tag{7}$$

Moreover, random genomes of length n have entropies differing from the upper bound $\lg_2(n)$ less than $\lg_2(n/(n - \lceil \lg_2(n) \rceil))$ (close to zero).

The proof states that due to Lemma 1, in the case $k = mrl(\mathbb{G}) + 1$ one has that $E_k(\mathbb{G}) = \lg_2(n-k+1)$. We argue that merely stating this does not prove eq. (7) Moreover, when stated in this form the result can be disproved when taking $k = 1$, $n = 4$ (for instance when $\mathbb{G} = ACTG$). In this case, one has:

$$E_k(\mathbb{G}) = \lg_2(n-k+1) = \lg_2(n) \not< \lg_2(n).$$

We suggest instead, to prove eq. (7) (in the correct range: $1 < k < n$) that one can simply use the fact that, for fixed n, the maximum number of different k-mers is $n-k+1$, and therefore the maximum entropy is obtained when they occur with the same probability $(n-k+1)^{-1}$, that is

$$E_k^{(\max)}(\mathbb{G}) = \sum_{i=1}^{n-k+1} \frac{1}{n-k+1} \lg_2(n-k+1) = \lg_2(n-k+1), \tag{8}$$

which implies

$$E_k(\mathbb{G}) \leq \lg_2(n-k+1) < \lg_2(n), \quad 1 < k < n. \tag{9}$$

Discussion

We show in this article that a number of mathematical statements presented in[5], on which the results of the paper rely, are, to our opinion, problematic. Some, although true, seem incorrectly proved while others appear false. We provide correct proofs for two of these results and show counter examples for the other. We would like to remark as well that the way in which these results are used to prove the main result of the article is also possibly incorrect but we fear it is too unclear for us to address it. Hence, we do not provide counter examples or corrections to this part. Last, we would like to note that the apparent agreement between the mathematical predictions and the empirical results of[5] could be justified, not by the correctness of their proofs, but by the decreasing importance of the possibility of overlapping k-mers when $n \gg k$.

Author contributions statement

M.A. and C.A. reviewed of the results in[5] and contributed to the proofs and corrections proposed, as well as the redaction of the manuscript. Both authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. S. Kurtz, A. Narechania, J.C. Stein and D. Ware. A new mwthod to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).
2. R. Hariharan, R. Simun, N. R. Pillai and T.D. Taylor. Comparative analysis of DNA word abundances in four yeast genomes using a novel statistical background model. *PLoS One* **8**, e58038 (2013).
3. M. Ghandi, D. Lee, M. Mohammed-Noori, M. A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol.* **10**, e1003711 (2014).
4. R. Wang, Y. Xu, B. Liu. Recombination spot identification based on gapped k-mers. *Sci. Rep.* **6**, 23934 (2016).
5. V. Bonnici and V. Manca. Informational laws of genome structures. *Sci. Rep.* **6**, 28840 (2016).
6. M. Li and P. Vitányi. *An introduction to Kolmogorov Complexity and its Applications* 3rd ed., Springer (2008).