Average (conditional) mutual information of protein coding and long noncoding RNA

John Napp[*]
Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Dated: May 18, 2018)

Information theoretic techniques have been used in bioinformatics to study a wide variety of statistical questions about various biological polymers. One commonly considered statistical property of a sequence (or family of sequences) is the *average mutual information*, a quantity inspired by the mutual information in information theory. Here, we define the *average conditional mutual information*, a quantity based on the conditional mutual information. While a small mutual information corresponds to approximately independent random variables, a small conditional mutual information corresponds to random variables which form an approximate Markov chain, so these measures capture quite different aspects of a probability distribution. We compute average mutual information and average conditional mutual information profiles for two types of transcripts obtained from the GENCODE database: (1) transcripts coding for proteins and (2) long non-coding RNAs, a class of transcripts which has seen a rise in interest in recent years. We discuss an exponential blowup issue associated with the average conditional mutual information which makes interpretation of the quantity difficult for transcripts of realistic length, and discuss possible ways to address this issue in the future. We find that protein coding transcripts, long non-coding RNAs, and randomly generated transcripts all have qualitatively different average (conditional) mutual information profiles.

I. INTRODUCTION

It is estimated that up to 70-90% of the human genome is transcribed at some point during development [1], although only a small fraction of these transcripts code for proteins. Researchers have long been trying to understand the biological significance of these non-coding transcripts. Do they serve a useful purpose or are they spurious? In this paper, we will focus on a particular class of non-coding transcript: long non-coding RNA (lncRNA). These are long RNA transcripts (generally considered to be of length at least 200 basepairs) which are not translated into proteins. We exclude from this definition well known functional RNAs such as ribosomal RNAs and transfer RNAs. Interest in lncRNA has risen sharply in the past decade, as researchers have implicated lncRNAs in certain diseases and developmental processes, and have pursued questions such as understanding how many types of lncRNAs there are, what (if any) the biological purpose of these molecules is, and how they function.

In this paper, we compare protein coding transcripts with lncRNAs using tools from information theory. This is similar in spirit to the 2000 paper [2], which used a quantity called the *average mutual information* (inspired by the mutual information from information theory) to compare coding versus noncoding DNA regions. They found that the average mutual information behaves quite differently in coding and noncoding regions. In this paper, we modify this analysis in two directions. First, instead of comparing protein coding transcripts against all noncoding DNA, we specialize to long non-coding RNAs. We find that the average mutual information profile for the lncRNA transcripts is qualitatively the same as what the authors of [2] found for (nonspecific) noncoding DNA.

The average mutual information has in fact been considered in many different contexts in bioinformatics [2–11]. Besides distinguishing coding from noncoding regions, it has also been used to detect correlated mutations [3] and as a genomic signature [11], as well as to study aspects of proteins (to name just a few). While the average mutual information is a statistical measure based on the information theoretic mutual information, to the best of my knowledge an analog of the information theoretic conditional mutual information has not been considered in this context. In this paper, we define such a quantity called the *average conditional mutual information.* We find that while the average conditional mutual information profiles of protein coding transcripts versus lncRNAs are qualitatively different, the average conditional mutual information suffers from an exponential blowup problem which makes interpretation problematic. In Section V, we discuss potential solutions to this problem as an avenue for future work.

II. PRELIMINARIES: MUTUAL INFORMATION AND CONDITIONAL MUTUAL INFORMATION

The mutual information is a measure of correlation between joint random variables. Precisely, given discrete random variables X and Y with joint probability distribution $p_{XY}(x,y)$, the mutual information $I(X:Y)$ between X and Y is given by

$$I(X:Y)_p = \sum_{x,y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \qquad (1)$$

where the logarithm (and all other logarithms in this paper) are base 2, and we take the convention that $0 \log \frac{0}{0} = 0$. Here, p_X denotes the marginal distribution

[*] napp@mit.edu

of X, and similarly for p_Y. The mutual information can also be expressed in terms of Shannon entropies:

$$I(X:Y) = H(X) + H(Y) - H(XY) \tag{2}$$

where $H(Z) := -\sum_z p(z)\log p(z)$ for any discrete random variable Z with probability distribution $p(z)$.

The mutual information enjoys a number of properties, including symmetry ($I(X:Y) = I(Y:X)$) and nonnegativity ($I(X:Y)_p \geq 0$ for any distribution p). Furthermore, $I(X:Y)_p = 0$ if and only if X and Y are independent. Recall that X and Y are independent if $p_{XY}(x,y) = p_X(x)p_Y(y)$. Finally, note that the maximal possible value of $I(X:Y)$ is given by $\min(\log|X|, \log|Y|)$, where $|X|$ denotes the number of possible values that the random variable X could take.

The mutual information was defined and an operational interpretation was given by Shannon in his seminal work [12], in which he related the mutual information to achievable communication rates over noisy channels. Intuitively, one may think of $I(X:Y)$ as a measure of correlation between X and Y, or as quantifying the amount of information about Y contained in X and *vice versa*. If X and Y are independent, $I(X:Y)$ is zero. As an example with high mutual information, consider the distribution given by $p_{XY}(0,0) = p_{XY}(1,1) = 1/2$. In this case, we have $I(X:Y)_p = 1$, and in fact this is the maximal possible mutual information when X and Y are binary random variables. Intuitively, $I(X:Y) = 1$ in this case because by learning X, one gains 1 bit of information about the value of Y.

We now turn to the conditional mutual information. Consider a probability distribution p_{XYZ} over three discrete random variables X, Y, Z. Then the mutual information between X and Z conditioned on Y, $I(X:Z|Y)$, is defined as

$$I(X:Z|Y)_{p_{XYZ}} = \mathbb{E}_Y I(X:Z|Y=y)_{p_{XZ|Y}}. \tag{3}$$

where $\mathbb{E}_Y$ denotes the expectation value over Y. In other words, $I(X:Z|Y)$ is the average of $I(X:Z|Y=y)$ over the possible values of Y. From this definition, it is clear that $I(X:Z|Y)$ is symmetric in the first two arguments and is nonnegative. Furthermore, it is true that $I(X:Z|Y)_{p_{XYZ}} = 0$ if and only if $X - Y - Z$ forms a Markov chain. That is, if and only if $p_{XYZ}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y)$. This characterization turns out to be robust, in the sense that if $I(X:Z|Y)$ is very small, then $X - Y - Z$ is very close to a Markov chain in a sense than can be made precise. Note that one can also express the conditional mutual information in terms of entropies:

$$I(X:Z|Y) = H(XY) + H(YZ) - H(Y) - H(XYZ). \tag{4}$$

It is important to note that given three random variables (X, Y, Z), one may have $I(X:Z|Y) < I(X:Z)$ or one may have $I(X:Z|Y) > I(X:Z)$. As an example in the former category, consider the distribution $p_{XYZ}(0,0,0) = p_{XYZ}(1,1,1) = 1/2$. In this case, $I(X:Z)_{p_{XZ}} = 1$, but $I(X:Z|Y)_{p_{XYZ}} = 0$ since Y completely determines the values of X and Z. As an example in the latter category, let $X \in \{0,1\}$ and $Z \in \{0,1\}$ be uniformly distributed, and let $Y = X + Z \mod 2$. In this example, $I(X:Z) = 0$ because X and Z are independent. But one may check that conditioning on Y induces a correlation between X and Z, and in fact $I(X:Z|Y) = 1$. These examples demonstrate that the conditional mutual information captures a much different aspect of a probability distribution than does the mutual information.

III. AVERAGE MUTUAL INFORMATION AND AVERAGE CONDITIONAL MUTUAL INFORMATION

We would like to apply the above information theoretic measures to study correlations in coding and lncRNA transcripts. To this end, we define two functions, the *average mutual information* (AMI) and the *average conditional mutual information* (ACMI), inspired by the mutual information and conditional mutual information quantities. The AMI has previously been considered in many different contexts in bioinformatics [2–11], but to the best of my knowledge the ACMI has not. We now define the AMI and ACMI functions associated with some nucleotide sequence T. Recall that T is a sequence of characters from the set $S := \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$.

First we define $AMI_T(k)$, the AMI function associated with transcript T. For each $x \in S$, let p_x denote the frequency of character x in T. For each pair $(x,y) \in S \times S$, let $p_{xy}(k)$ denote the frequency of (x,y) amongst all pairs of characters a distance k apart in T (i.e., separated by $k-1$ characters). We now define the average mutual information as

$$AMI_T(k) := \sum_{(x,y)\in S\times S} p_{xy}(k) \log \frac{p_{xy}(k)}{p_x p_y}. \tag{5}$$

It is clear by comparison with the expression for mutual information in Section II that the two quantities are related. For example, if there are no correlations in the sequence, we expect $p_{xy}(k) \approx p_x p_y$ for all $(x,y) \in S \times S$, and we would obtain $AMI_T(k) \approx 0$. But it is important to note that the quantities are also fundamentally different. The mutual information between two random variables is determined by their joint distribution. But a fixed transcript T is a deterministic quantity. $AMI_T(k)$ is in some sense considering all pairs of characters a distance k apart as a set of samples coming from the same joint distribution, and empirically estimating the mutual information for this joint distribution. Clearly this interpretation should not be taken too literally, but one can simply think of AMI as a function defined for a transcript T which is inspired by the mutual information. One issue to keep in mind is that $AMI_T(k)$ is better behaved for longer transcripts T. The smaller T is, the more sensitive $AMI_T(k)$ is to small changes in the sequence. For example, a randomly generated transcript of length 10 could easily have a very large AMI, but a randomly generated transcript of length (say) 3000 will have a very small AMI with overwhelming probability as a result of the central limit theorem. For this reason, in the analysis below we will only consider long transcripts (of length at least 3000).

We now define the average conditional mutual information (ACMI) function associated with some transcript T. For each $z \in S^{\times(k-1)}$ (that is, for each length-$(k-1)$ subsequence), define q_z to be the frequency of subsequence z amongst all subsequences in T of length $k-1$. For $x \in S$, define $l_{x|z}$ to be the frequency of character x, amongst all characters appearing to the left of a z subsequence in T. Similarly, define $r_{y|z}$ to be the frequency of character y, amongst all characters appearing to the right of a z subsequence in T. Finally, amongst all occurrences of a z subsequence in T, define $p_{xy|z}$ for $x, y \in S \times S$ to be the frequency of a xzy subsequence. As a technical detail, note that when calculating the q_z, we do not count subsequences which contain the first or last character.

We illustrate these definitions with a simple example. Take $T = \mathtt{AGCCTGCA}$. Then we have $q_A = 0$, $q_C = 1/2$, $q_G = 1/3$, $q_T = 1/6$, $q_{CC} = 1/5$, $q_{CT} = 1/5$, $q_{GC} = 2/5$, and $q_{TG} = 1/5$. We also have, for example, $l_{A|GC} = l_{T|GC} = 1/2$ and $r_{A|GC} = r_{C|GC} = 1/2$. We also have $p_{AC|GC} = p_{TA|GC} = 1/2$.

We are now ready to define $ACMI_T(k)$ for any sequence T:

$$ACMI_T(k) := \sum_{z \in S^{\times(k-1)}} q(z) \left(\sum_{(x,y) \in S \times S} p_{xy|z}(k) \log \frac{p_{xy|z}(k)}{l_{x|z} r_{y|z}} \right). \tag{6}$$

Note the strong resemblance to the conditional mutual information defined in Eq. 3.

A. Examples

We now provide some numerical examples to illustrate and benchmark the above functions. Our first example is a uniformly random sequence. The second is a completely deterministic sequence. The third is a sequence that is artificially constructed to have large $ACMI(2)$ but small AMI.

1. *Uniformly random sequence*

We generated a uniformly random sequence of length 3000 (Figs. 1, 2). We found that $AMI(k)$ is close to zero. This is expected, as a uniformly random sequence has no correlations. However, we find that $ACMI(k)$ is small but nonzero for $k = 2$ and $k = 3$, but becomes larger as k is increased. This is due to an exponential blowup issue which ACMI has but AMI does not. In particular, when calculating $ACMI(k)$, we must consider conditioning on all possible subsequences of length $k-1$. But this number of subsequences grows as 4^{k-1}, exponentially quickly as a function of k. This is both a computational and statistical bottleneck. Computationally, the time complexity of computing $ACMI_T(k)$ for some transcript T is exponential in k. Statistically, we expect that $ACMI_T(k)$ is meaningful when $4^{k-1} \ll |T|$ where $|T|$ is the length of T. Already when $k = 6$, we have $4^{6-1} = 1024$. This explains why $ACMI(k)$ is actually quite large for $k = 6$ in the uniformly random case of length 3000. Indeed, if we consider a uniformly random sequence of length 200000, we find that $ACMI(k)$ is very small even for $k = 6$.

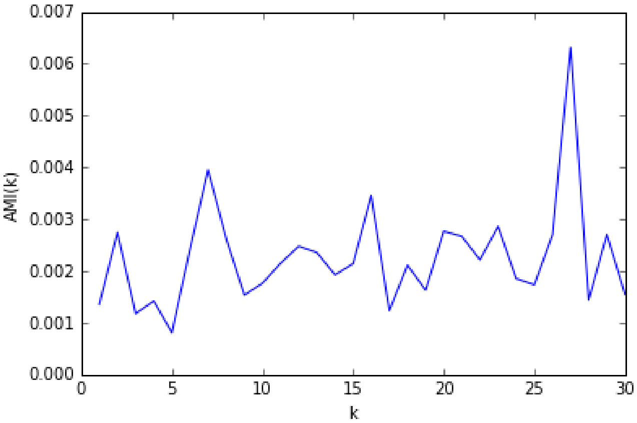

FIG. 1. AMI profile of a uniformly random transcript of length 3000. Note that $AMI(k)$ is close to zero for all k.

2. *Deterministic sequence*

We generated a sequence of the form $\mathtt{ACGTACGTACGTACGTACGT...}$ of length 3000 (Figs. 3, 4). We found that $AMI(k)$ is extremely close to 2 for all k, and $ACMI(k)$ is extremely close to 0 for all k. This is as expected. Indeed, since the sequence is deterministic, learning any particular bit gives you 2 bits of information about the bit a distance k away, so $AMI(k) = 2$. But since conditioning on the value of any particular bit gives one complete information about the

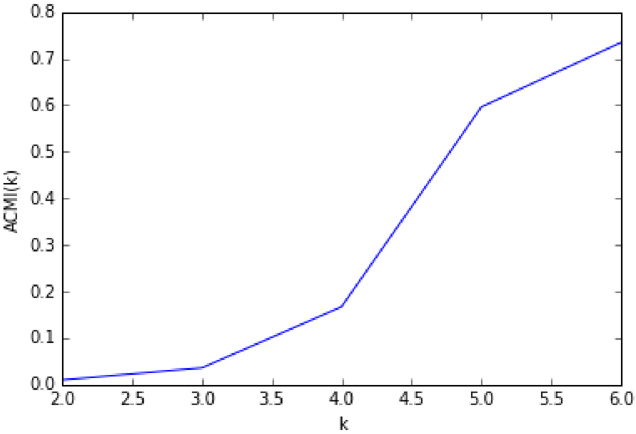

FIG. 2. ACMI profile of a uniformly random transcript of length 3000. $ACMI(k)$ is small for $k = 2$ and $k = 3$, but becomes substantial as k grows. This is due to finite-size effects and an exponential blowup issue of ACMI, as discussed in the main text.

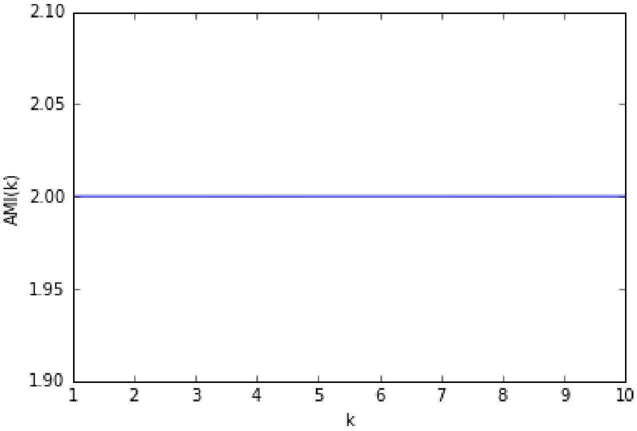

FIG. 3. AMI profile of the transcript ACGTACGT... of length 3000. Note that $AMI(k)$ is 2 for all k, since learning one character of the sequence gives 2 bits of information about the character distance k away.

entire sequence, $ACMI(k) = 0$ for all k.

3. *High $ACMI(2)$, low AMI sequence*

Finally, we generate a sequence which is designed to have high ACMI (for $k = 2$) but small AMI (Figs. 5, 6). To do this, we uniformly at random generate a sequence consisting of integers $\{0, 1, 2, 3\}$. Next, between each pair of adjacent integers in this sequence, we insert an integer between them equal to their sum modulo four. We then convert the sequence of integers into a sequence of A,C,G,T in some canonical way. We created a sequence of length 3000 in this way. Note that characters in the resulting sequence are pairwise independent, so $AMI(k)$ is very small as expected, but conditioning on the value of character j may induce a strong correlation between characters $j - 1$ and $j + 1$, so $ACMI(k)$ is indeed relatively large for $k = 2$.

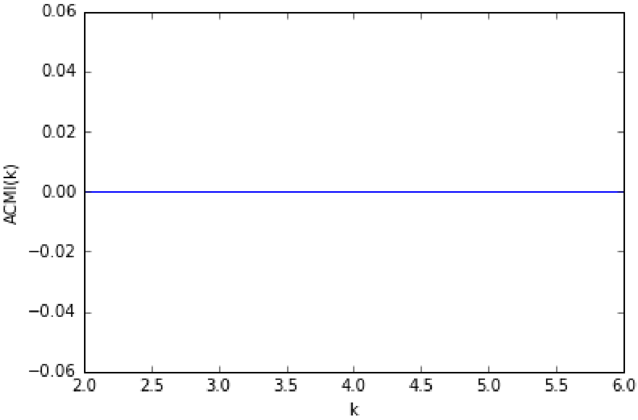

FIG. 4. ACMI profile of the transcript ACGTACGT... of length 3000. Note that $ACMI(k)$ is zero for all k.

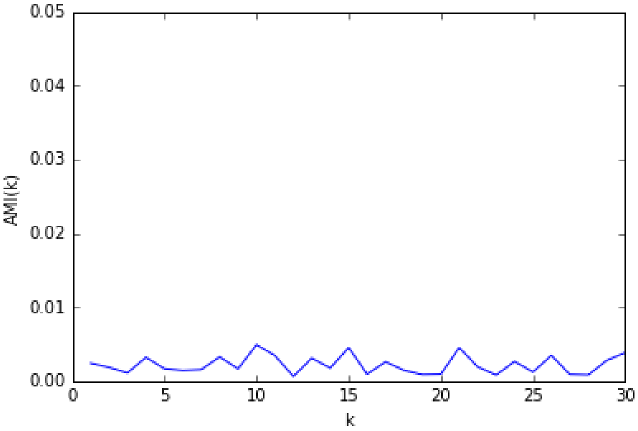

FIG. 5. AMI profile of the sequence designed to have high $ACMI(2)$ and low AMI as discussed in the main text. Note that $AMI(k)$ is nearly zero for all k.

IV. AMI AND ACMI OF PROTEIN CODING AND LONG NON-CODING RNA TRANSCRIPTS

In this section, we apply the statistical techniques defined above to study the AMI and ACMI profiles of two types of transcripts: (1) transcripts which code for proteins, and (2) long non-coding RNA (lncRNA) transcripts. Note that a similar analysis was performed in [2]. In that paper, the authors considered the AMI function, and applied it to coding and non-coding DNA. They found that the AMI function had different profiles for coding versus non-coding regions. Our analysis differs from theirs in two respects. First, we specialize the non-

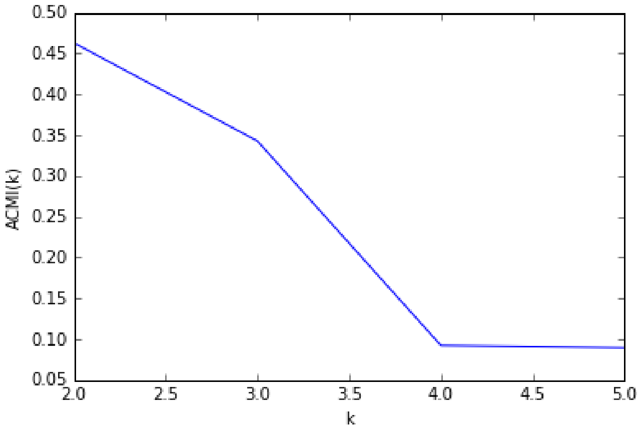

FIG. 6. ACMI profile of the sequence designed to have high $ACMI(2)$ and low AMI as discussed in the main text. Note that $ACMI(k)$ is very high for this sequence for $k = 2$.

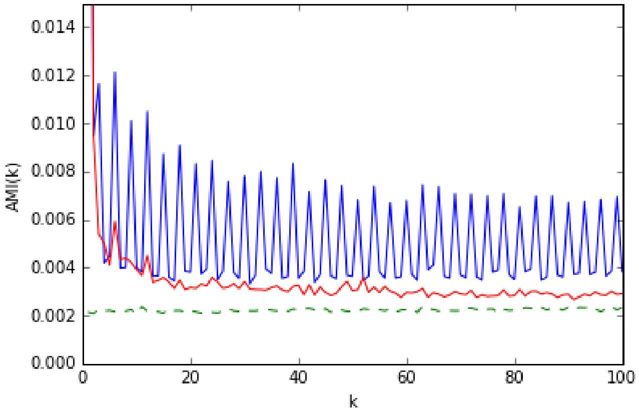

FIG. 7. AMI profile of human **(blue)** protein coding transcripts, **(red)** lncRNA transcripts, **(green)** uniformly random transcripts. Each plot is an average over 300 transcripts. Only transcripts of length at least 3000 were considered, and the lengths were truncated to exactly 3000 to mitigate finite-size biases.

coding DNA regions we study to those which correspond to lncRNA. Since many lncRNA transcripts are believed to possibly be functional, it is interesting to specialize to this case and see if the AMI profile still looks very different than that of protein coding DNA. Second, we compute not only the AMI profiles, but also the ACMI profiles.

Specifically, we analyzed [13] human protein coding and lncRNA transcripts obtained via the GENCODE project [14] database. Of the transcripts available, we randomly selected 300 protein coding sequences and 300 lncRNA sequences of length at least 3000. To mitigate finite-size biases, we then truncated the sequences to length exactly 3000. For each sequence, $AMI(k)$ was computed for $k = 1 \ldots 100$ and $ACMI(k)$ was computed for $k = 2 \ldots 6$. Finally, for both sets of transcripts, the 300 $AMI(k)$ profiles and the 300 $ACMI(k)$ profiles were averaged to obtain an average AMI and average ACMI profile for both sets of transcripts.

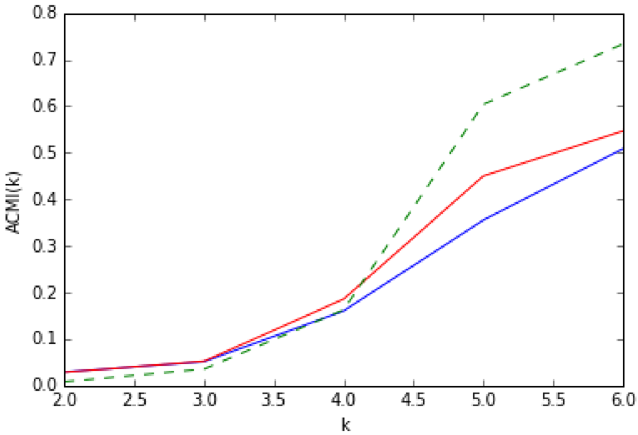

FIG. 8. ACMI profile of human **(blue)** protein coding transcripts, **(red)** lncRNA transcripts, **(green)** uniformly random transcripts. Each plot is an average over 300 transcripts. Only transcripts of length at least 3000 were considered, and the lengths were truncated to exactly 3000 to mitigate finite-size biases.

V. DISCUSSION

Consider first the AMI profiles of the coding and lncRNA transcripts (Figs. 7). We see that the AMI profile for the coding transcripts has small spikes occurring every three nucleotides. This is qualitatively consistent with the results of a similar analysis in [2]. The AMI we have found in this paper is larger than the AMI found in that paper, but this is not necessarily an inconsistency. They used sequences of length 500, and we used sequences of length 3000. They also did additional processing to attempt to mitigate the finite-size effect, whereas we did not. Further analysis would be required to determine if the two results are consistent.

The periodic structure of the AMI for coding transcripts can be understood as arising from the genetic code. Codons have length three, and different codons appear with different frequencies in the genome. From this perspective, it's not surprising that the *in-frame* AMI (i.e., $k = 3n + 1$ for $n \in \mathbb{Z}_+$) should be higher than the *out-of-frame* AMI. We find that the AMI profile for lncRNA decays quickly to a very small value, but it is interesting to note that even at $k = 100$, the AMI for lncRNA is greater than that of randomly generated transcripts, which hovers around 0.002 due to finite-size effects.

We now consider the ACMI profile for protein coding versus lncRNA transcripts (Fig. 8). We find that the ACMI profiles for protein coding, lncRNA, and uniformly random transcripts are all qualitatively different from each other. Due to the exponential blowup issue associated with computing the ACMI discussed above, we only compute $ACMI(k)$ for $k \leq 6$. Note that one should not try to seriously interpret Fig. 8 in terms of conditional mutual information, since the data seems to be dominated by finite-size effects. In particular, for $k = 5$ and $k = 6$, the ACMI for the uniformly random case is quite large, while the ACMI for the uniformly random case should actually approach zero in the asymptotic limit. It is perhaps somewhat interesting that the ACMI profiles for the protein coding and lncRNA cases are qualitatively fairly similar, while for the case of AMI they look drastically different.

This line of work could be continued along several lines. For one, a more detailed analysis could be carried out to study the impact of finite-size effects, and the AMI and ACMI functions could be defined in more complicated manners to attempt to mitigate these biases. It was somewhat disappointing that the ACMI profiles were dominated by finite-size effects, but not very surprising given the exponential blowup associated with the ACMI definition. Recall that this blowup occurs when calculating $ACMI(k)$ because we need to consider separately conditioning on every possible subsequence of length $k-1$, which is 4^{k-1} subsequences. Another avenue for future work could be to attempt to alleviate this issue by introducing a coarse-graining, so that instead of separately conditioning on 4^{k-1} possibilities, one conditions on a much smaller number of possibilities. In this manner, one could obtain better statistics while sacrificing some granularity.

VI. CONCLUSION

We introduced and defined the average conditional mutual information (ACMI) as a statistical tool to study correlations in DNA sequences. The ACMI is related but distinct from the previously considered average mutual information (AMI). As a benchmark, we computed AMI and ACMI profiles for three types of artificial sequences of length 3000 base pairs: a uniformly random sequence, a deterministic sequence ACGTACGTACGT..., and a sequence designed to have high ACMI for $k = 2$. The relatively large ACMI values for the uniformly random sequence demonstrated the significant finite-size effects for the ACMI function, which arise due to an exponential blowup issue in the definition of the quantity.

We computed the AMI and ACMI profiles for (1) human protein coding transcripts and (2) long noncoding RNA transcripts, obtained from the GENCODE database. Comparing our results for the AMI profiles to those of [2] suggested that with respect to AMI, lncRNA is not much different than nonspecific noncoding DNA. The ACMI profiles of protein coding, lncRNA, and random transcripts were all qualitatively different from each other, although they were more similar to each other than they were for the AMI case. However, to be a useful tool in bioinformatics, the exponential blowup issue of the ACMI should be addressed, which could be a direction for future work.

[1] J. T. Kung, D. Colognori, and J. T. Lee, Genetics **193**, 651 (2013).

[2] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, Physical Review E **61**, 5624 (2000).

[3] B. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes, Proceedings of the National Academy of Sciences **90**, 7176 (1993).

[4] I. L. Hofacker, M. Fekete, and P. F. Stadler, Journal of molecular biology **319**, 1059 (2002).

[5] S. Lindgreen, P. P. Gardner, and A. Krogh, Bioinformatics **22**, 2988 (2006).

[6] R. Roman-Roldan, P. Bernaola-Galvan, and J. Oliver, Pattern recognition **29**, 1187 (1996).

[7] B. Giraud, A. Lapedes, and L. C. Liu, Physical Review E **58**, 6312 (1998).

[8] H. Herzel and I. Große, Physical Review E **55**, 800 (1997).

[9] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, Proceedings of the National Academy of Sciences of the United States of America **102**, 18297 (2005).

[10] N. Slonim, O. Elemento, and S. Tavazoie, Molecular systems biology **2** (2006).

[11] M. Bauer, S. M. Schuster, and K. Sayood, BMC bioinformatics **9**, 48 (2008).

[12] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[13] Code available upon request.

[14] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, *et al.*, Genome research **22**, 1760 (2012).