RESEARCH

Mutual Information Reveals a 6 Base Pair Signal in Noncoding DNA

James L. Dannemiller

Correspondence:
dannemil@rice.edu
Department of Psychological Sciences MS-25, Rice University, 1600 Main St., 77005 Houston, USA
Manuscript posted: October 1, 2019
Full list of author information is available at the end of the article

Abstract

Background: Coding DNA in eukaryotes shows a prominent periodicity at 3–base pair (bp) because of the nonuniform use of bases at the three codon positions. This periodicity can also be interpreted as a correlation between bases at a distance of 3–bp, which in turn implies enhanced *predictability* at that distance as well as at integer multiples of 3 bp. This study examined possible similar dependencies between nucleotide positions in human promoter region sequences. It is possible that such noncoding sequences might show enhanced predictability at various distances other than 3 bps because of the ubiquitousness of repetitive patterns such as Short Tandem Repeats (STR). It is also possible that the $3D$ conformational requirements of helical DNA could induce a dependency between bps occupying certain relative positions. Mutual Information (MI), an information–theoretic quantity that can be interpreted as a measure of correlation or predictability between two sequence positions, was used to probe promoter sequences in *Homo sapiens*. Coding DNA sequences were also probed as a methodological control.

Results: The MI measure showed a clear 3–bp signal with integer harmonics in coding DNA sequences. This shows that the MI measure is sensitive to a known repetitive signal in coding DNA sequences. MI in promoter sequences fell with increasing distance between two positions, but also showed an unexpected and significant rise at a distance of 6–bp (i.e., between the first and the seventh bases in a sequence). Harmonic intrusion of a 3–bp signal from coding DNA was ruled out as an explanation for this 6–bp signal in promoter DNA. Purposely misaligning the promoter sequences had essentially no effect on the increased correlation at a distance of 6 bp.

Conclusions: Human promoter DNA sequences shows enhanced predictability (i.e., increased mutual information) between a given position and a position 6 bp downstream. This signal runs counter to the general fall in mutual information as the distance between positions increases. The increased correlation at a distance of 6 bp does not depend on the precise alignment of the sequences. Possible explanations for this enhanced predictability are discussed.

Keywords: mutual information; noncoding; promoter; sequence predictability

Background

Dependencies between bases separated by various distances have been documented in coding DNA [1]. Coding DNA sequences (CDS sequences) tends to show a 3–bp periodicity because of the nonuniform use of the four bp at different codon positions [2, 3]. Longer range correlations have also been reported both for prokaryotic and eukaryotic genomes [4–6]. Intronic sequences in human DNA also show long–range correlations [7].

Dependencies between sequence positions outside of CDS sequences or introns could exist for several reasons. Short Tandem Repeats (STR) could produce repetitive sequences with a periodicity that depends on the length and composition of the repeating unit. Promoter regions contain an abundance of STR [8]. An information–theoretic analysis showed that the entropies of various genomic sequences depended on their proportions of repeats [9]. Indeed, it has been shown that there are more pairs of identical single nucleotide polymorphisms (SNP) within STR separated by 2, 4, 6 and 8 bp than by 1, 3, 5, 7 or 9 bp, and that this pattern is observed mainly within STR [10].

Another reason for possible dependencies in short DNA sequences could be the 3–D helical structure of DNA. DNA in its B conformation has a helical repeat cycle of approximately 10 to 10.5 bp [11, 12]. The curvature of DNA has also been shown to be sequence–dependent [13]. It is difficult, however, to make the case that bp distances in 3–D space related to the helical structure of DNA strongly constrain the presence of particular nucleotides at various positions. If this were the case, then those constraints would be expected to operate more or less uniformly across the genome, and as noted above, periodicities most likely differ between coding and noncoding regions. Differences in chromatin structure between regions could still make it possible that the distances in 3–space could play some role in constraining base usage at certain relative positions.

Uncovering correlations between bp separated by specific distances is difficult because the use of a standard Pearson's r correlation coefficient is inappropriate. Pearson's r correlation coefficient is intended to measure the correlation between numeric random variables. The four DNA bases are instead nominal or symbolic. It is inappropriate simply to recode the four DNA bases numerically (e.g., 1, 2, 3 and 4) and then compute Pearson's r because that converts a nominal Random Variable (RV) into one that is numeric in appearance only.

An alternative measure is Mutual Information (MI) (see Methods). MI is an information-theoretic measure that can be calculated over jointly distributed symbolic sequences. MI also captures *any* dependency, not just linear, between random variables. For example, across a large number of aligned DNA sequences, if the base A appearing at position x tended to be followed more often than not by the base T at position x + 5, this would result in a significant value of the MI coefficient at a distance of 5 bp. Similarly, if a particular region within these aligned sequences tended to contain many monomeric repeats, MI would also signal this nonrandom pattern.

MI has been used previously to reveal short range dependencies in both coding and noncoding DNA. There is a prominent repetitive MI signal at 3 bp in human protein–coding DNA sequences [2]. The period of 3 bases is most likely the result of the four bases being used nonrandomly across the three codon positions. That periodic signal was absent in another study that used noncoding sequences [2]. The authors of this study noted that MI in noncoding DNA was statistically above chance levels at short distances between positions but fell quickly toward chance levels as distance increased. It should be noted, however, that the function relating MI to distance between bp did not fall smoothly with increasing

distance. A closer inspection of Figure 1 in [2] shows a clear spike in the value of MI at a distance of 6–bp in noncoding DNA. The authors did not address this spike in MI at 6–bp choosing instead to focus on the general pattern of MI falling with distance in noncoding DNA.

The purpose of the present study was to examine more thoroughly the possibility of a correlation between the bases in human promoter DNA sequences that occupy positions at l and $l+6$. Such a result appears to be present in [2]. To accomplish this, MI was calculated at distance from 1 to 21 bp for over 22,000 human promoter sequences aligned at their putative Transcription Start Site (TSS).

Results

The results using the bias–corrected Mutual Information (bcMI) measure with the CDS sequences separately by strand are shown in Figure 1. With the exception of some minor differences at distances of 2 and 3–bp, the results are virtually identical out to a distance of 21 bp for sequences on both strands. There are very prominent spikes in bcMI at distances that are even and odd harmonics of 3–bp (6 , 9, 12 bp etc.). In addition, the largest bcMI value is observed on both strands at a distance of 1 bp. In other words, the correlation is strongest between bases at adjacent positions.

Turning next to the results with promoter sequences, bcMI vs. distance (in bp) was computed separately for the + and - strand sequences using various starting positions in the range from -1000 to -50 bp re: TSS (re: TSS). Figure 2 shows bcMI plotted against distance separately for plus strand sequences and minus strand sequences. The error bars are $\pm 2SEM$. The general features are the same for both subsets of sequences. bcMI falls as the distance from the first to the second position increases. This confirms the result shown in [2]. There is, however, a conspicuous and significant increase in bcMI at a distance of 6–bp for both subsets of sequences. This plot shows that the predictability of a downstream base from a fixed upstream base is significantly higher when that base is 6–bp downstream than when it is 3, 4, 5, 7, 8, 9 … 21 bp downstream.

The data in Figure 2 were generated by computing bcMI for all possible pairs of positions from -99 to -57 bp re: TSS. In other words, the starting position was set to -99 bp re: TSS, and bcMI was computed for all distances from 1 to 42 bp (i.e., positions -98 to -57 bp re: TSS). Recall, that each time bcMI is calculated, the simple and joint probabilities are estimated from the relevant proportions of bases across all 11,602 (plus strand) or 11,117 (minus strand) sequences. This procedure produces a 42 x 42 matrix of bcMI values. The major diagonal holds the bcMI values between a starting position from -99 to -57 bp re: TSS and itself (mutual self–information). Once these are diagonal entries are eliminated, consider the remaining upper triangle of this matrix. The first row of the upper triangular matrix holds the bcMI values between the starting position of -99 bp re: TSS and second positions from -98 to -57 bp re: TSS: in other words, for distance from 1 to 42 bp. The second row holds the bcMI values between a starting position of -98 bp re: TSS and second positions from -97 to - 57: distance from 1 to 41 bp. Continuing in this manner, the 21st row holds the bcMI values for a range of distances from 1 to 21 bp.

Notice, then, that the +1 diagonal in the upper triangular matrix holds exclusively bcMI values for distances of 1 bp; that is, between positions -99 and -98, -98 and -97, -97 and -96, etc. The + 2 diagonal holds bcMI values exclusively for distances of 2. Traversing this upper triangular matrix to the + 21 diagonal gives 21 estimates of bcMI for a distance of 21 bp.

To produce the data points in Figure 2, the first 21 values of each successive diagonal from + 1 to + 21 were averaged. Thus, in Figure 2 the data point for the plus strand sequences at a distance of 1 bp is the average of 21 bcMI values calculated between positions -99 and -98, -98 and -97, 97 and -96, . . ., -78 and -77. Each data point in Figure 2 is the average of 21 bcMI estimates using -99 bp re: TSS as the initial position.

Figure 3 is a visualization of one of these upper triangular portions of a square bcMI matrix. In this case, only the plus strand sequences were used. bcMI was computed for all possible pairs of positions between -99 and 0 bp re: TSS. This is the analogue of a correlation matrix with bcMI values instead of Pearson r correlation estimates populating the matrix. The color saturation of a cell is proportional to the bcMI value between those two positions. The leftmost, longest diagonal represents estimates of bcMI for a distance of 1. The bcMI values in the matrix have all been raised to the 0.27^{th} power (empirically determined) to make it easier to see the bcMI values that are less than those along the first diagonal. Without this monotonic transformation, the matrix is dominated by the distance−1 diagonal. There is clearly a signal at a distance of 6−bp (the + 6 diagonal is indicated by red line). Although there is variance across the various starting positions from upper left to lower right along this +6−bp diagonal, the signal at a distance of 6−bp is present across this entire range from -99 to the TSS. There is some weak evidence for a signal at a distance of 12 bp, especially for the starting positions closer to the TSS.

To examine how the magnitude of this 'spike' in mutual information varies with distance to the putative TSS, bcMI was calculated using the plus strand sequences for 122 discrete starting positions from 1000 bp upstream of the TSS to 574 bp downstream of the TSS in steps of 13 bp. bcMI was calculated for distances from 1 to 12 bp. Figure 4 shows these 122 curves. The lowest curve has a starting position (0 on the x−axis) that is 1000 bp upstream of the putative TSSs. For each subsequent curve, the inititial nucleotide (*nt*) steps toward to TSS by 13 bp. Additionally, each subsequent curve has been shifted vertically by 0.00005 MI units so the relative values of MI within a curve have been preserved, but the absolute values of MI have not. The rise in MI at a distance of 6−bp is a prominent feature throughout the promoter range from 1000 bp upstream of the TSS to 574 bp downstream of the TSS. There is some indication that the magnitude of this spike in MI increases as the initial nucleotide approaches and moves downstream of the TSS. There are also weaker but evident spikes at distances of 9 and 12 bp, especially for starting positions downstream of the TSS.

These trends can be seen more clearly in Figure 5. In this figure, the bcMI values at distances 5, 6 and 7 bp shown in Figure 4 have been plotted in separate panels to show the relative magnitude of the spike at a distance of 6−bp. For all three of these distances, there is a rise in MI as the TSS is approached with MI at a distance of 6−bp being higher than at 5 and 7 bp. At all three distances, there is an abrupt drop in MI at the putative TSS. This is followed by a gradual rise in MI through positions approximately 250 −300 bp downstream of the TSS, and finally a gradual fall in MI out to 574 bp downstream of the TSS.

This last trend is worth considering. If a signal at a distance of 3−bp from the intrusion of protein −coding DNA were responsible for the spike at a distance of 6−bp seen at the downstream starting positions in Figure 4, then one would expect to see MI values downstream of the TSS at a distance of 6−bp rise and stay high for starting positions farther downstream from the TSS. But the middle panel in Figure 5 shows that following an initial rise immediately after the TSS for a distance of 6−bp, MI falls as the starting position moves farther downstream of the TSS until it is clearly below the values just upstream of the TSS. This

argues that the spike in MI at a distance of 6–bp is most likely not the result of intrusion of parts of CDS sequences into the MI calculation for promoter sequences.

Additionally, unlike MI for the CDS sequences sequences shown in Figure 1 where there is a prominent signal at a distance of 3 bp, the MI signal at 3 bp in the promoter sequences as shown in Figure 2 is not nearly as strong as the 6 –bp signal. In fact, the MI signal at 3 bp in promoter sequences is weaker than the signals at distances of 2 and 4 bp.

To determine whether more global aspects of these promoter sequences might play a role in producing the features of MI versus distance, the sequences coded on the plus strands were split into two groups based on G+C content over the range from -300 to -100 bp re: TSS. MI was calculated over the 22 positions just 3′ of this region (i.e., -99 to -78 bp re: TSS) with -99 bp re: TSS as the fixed initial position. Figure 6 shows the results. Two points are worth noting. First, the rise at a distance of 6 bp appears to be larger in the sequences with higher G+C (G+C) content. This 6–bp signal is present, but much weaker in the sequences with lower G+C content (higher A+T content). Second, both the sequences with high and low G+C content show their strongest correlations between adjacent positions as was evident in the complete set of sequences.

In the sequences classified as high *vs.* low G+C content, the most frequent heptamers were determined. Shown highlighted in red and labeled in Figure 7 are any heptamers that reached a frequency of at least 10 tallied across all of the sequences (see caption to Figure 7 for the interpretation of the grids in these panels). Heptamers were examined because of the possibility that the first base in the sequence and the seventh base in the sequence were identical. The seventh base in a heptamer sits 6 bp downstream of the initial base. Eight of the ten more frequent hexamers appeared in the high G+C content sequences. Of these eight hexamers (CCCCGCC, GGGAGGG, GCCCCGC, GGGCGGG, CCCGCCC, CCGCCGC, CGCCGCC, GGCGGGG, CCAGAGA, and CATGCAA), seven of them repeat the initial base in the seventh position, while neither of the more frequent hexamers in the low G+C content sequences have this property. These results with the high *vs.* low G+C content sequences are discussed further below.

Finally, the results in Figure 4 suggest that the spike in MI at a distance of 6 bp is characteristic of human promoter sequences throughout the range from 1000 bp upstream of the TSS to 500 bp downstream of the TSS. This result leads to the following prediction:

> *The function relating MI to distance should be approximately independent of the alignment of the sequences.*

The logic of this prediction is that the curve relating MI to distance is determined by the correlation between bases at positions *within* sequences. Examining Figure 4 shows that if half of the sequences were used to compute MI using one starting position, and the other half of the sequences were used to compute MI at a second starting position, the resulting MI *versus* (*vs.*) distance curve should be the average of the two curves. Following this logic, it should be possible to subdivide the set of N sequences into approximately equally –sized subsets (e.g. sequences $(1\ldots j)$, $(j+1\ldots 2j)$, …, $(N-j\ldots N)$, compute MI *vs.* distance for each subset, and average these curves. The result should look approximately like one of the MI *vs.* distance curves in Figure 4.

If this logic is followed to its extreme, it should be possible to compute MI *vs.* distance using a different starting position for each sequence (i.e., subdivide the set of sequences into N subsets each of size 1), and the resulting curve computed over all of the misaligned

sequences should look approximately like the curves in Figure 4. To test this hypothesis, a starting position was selected as -200 bp re: TSS for the aligned sequences. To create the set of misaligned sequences, a random integer step in the range of uniformly distributed integers from ± 40 bp was generated for each sequence, and the sequence was shifted by that step such that its original starting position of -200 bp re: TSS was replaced by its shifted starting position. For example, if the random steps were -24 and + 10 bp for two given sequences, then one sequence would be shifted by 24 bp in the $5'$ direction and the other by 10 bp in the $3'$ direction. Whereas before the shift, each sequence used a starting position of -200 bp for the computation of MI *vs.* distance, now these two sequences would have MI computed starting at -224 and -190 bp re: TSS. It is important to note that this misalignment manipulation preserves the relative positions of the bases *within* a sequence.

The result of this purposeful misalignment of the sequences is shown in Figure 8. As predicted, there is essentially no effect of computing MI on sequences that are misaligned within the range of ± 40 bp.

Discussion

MI was used to assess dependencies between positions in human promoter sequences. First, however, it was important to establish that this information–theoretic measure was sensitive to known dependencies across DNA positions. For this reason, the methodological control involved the use of CDS sequences. In past research, analysis of dependencies between positions in these sequences has shown a strong signal at 3 bp as well as at the integer harmonics of 6, 9, 12, bp etc. These same results were replicated here using the MI measure of correlation between positions. There was a clear correlation between bases at a distance of 3 bp in CDS sequences. The 3–bp signal could be the result of unequal base occupancy of codon positions in CDS sequences. For example, the third letter in four-degenerate synonymous-codon amino acids shows biases in base usage [14]. Such biases would make the bases that occupied the third letter in such codons more predictable than if there were uniform usage of the four bases independently at this third position. This, in turn, would result in higher MI values for positions that differed by multiples of 3–bp, especially to the extent that multiple copies of the same amino acid followed each other in the resulting protein sequence. These results with CDS sequences show that the MI measure used in the present study is sensitive to known statistical dependencies between positions in CDS sequences.

In the human promoter sequences, there were two notable results. First, as in the CDS sequences the largest mean MI values occurred between adjacent positions. Second, there was a significant out–of–trend rise in MI at a distance of 6 bp. This 6–bp signal was present in sequences specified on both strands. As was noted above, such a result is evident in Figure 1 of [2]. The authors of that study did not discussa nor draw attention to this rise in MI at a distance of 6 bp perhaps because MI fluctuated quite substantially over many of the shorter distances. The authors instead chose to focus on the overall trend of MI decreasing with increasing distance. The same overall trend was evident in the present results, but in addition, a clear signal at a distance of 6 bp was also present. These two results will be discussed in turn.

As in CDS sequences, the strongest correlation in promoter sequences was between adjacent positions. Why would MI values be strongest in promoter sequences at adjacent positions? One explanation could be that certain more global features of promoter sequences lead to this result. Recall, that in Figure 6, sequences with higher *vs.* lower $G+C$ content both

showed their highest mean MI values at adjacent positions. One might have expected to see a somewhat higher mean MI value for the high $G+C$ content sequences by virtue of the fact that such sequences probably also exhibit more CC, CpG, GG and GpC dinucleotide repeats in this region. In fact, the mean MI at adjacent positions was slightly higher in the low $G+C$ content sequences.

The results based on the $G+C$ content split were more interesting for the rise in mean MI at a distance of 6 bp. The magnitude of mean MI and the increase in MI at a distance of 6 bp relative to distances slightly smaller or larger was found found to be larger in the sequences with higher $G+C$ content. In the sequences with lower $G+C$ content, there was weaker evidence for this 6–bp MI signal. An analysis of the more frequent heptamers that appeared across the range from -300 to -100 bp re: TSS showed that of the 10 most frequent heptamers, eight of them occurred in the sequences with high $G+C$ content. Of these eight heptamers (see Results for a listing), seven of them showed the same nucleotide in the first and seventh positions; that is, the base at position n was repeated 6 bp downstream. These repeated bases at a distance of 6 bp would tend to increase the value of MI at that distance.

Finally, the manipulation that purposely misaligned the sequences before computing MI showed essentially no effect on the MI *vs.* distance curve. In particular, the spike at a distance of 6 bp was still present. The resulting MI *vs.* distance curve can be thought of as approximating the average of the 80 curves with starting positions in the range from -200 $\pm$ 40 bp generated using the aligned sequences. This result also argues that the increased correlation at a distance of 6 bp is a property of promoter sequences that is not tied to one specific region of the sequence. Rather, statistically, it characterizes each sequence throughout its range from -240 to -160 bp re: TSS and most likely throughout the entire range from 1000 bp upstream to approximately 500 bp downstream of the TSS. As noted above, there is variation in the strength of this distance–6 bp effect that depends on the $G+C$ content of the sequence.

Conclusions

Human promoter DNA sequences show a prominent correlation between bases at positions d and $d+6$ bp throughout a range at least 1000 bp above the TSS. Mutual information was used to probe these correlations. This correlation at a distance of 6 bp is significantly higher than the correlations at distances of 3 to 5 bp and 7 to 21 bp. The correlation generally falls with increasing distance from its highest value for neighboring bases, but the fall is interrupted by the rise in the correlation at a distance of 6 bp. This aspect of promoter sequence statistics has been seen before in published analyses, but it was never discussed or followed up in that previous work [2]. The possibility that this represents a harmonic intrusion of a distance 3 bp signal in CDS sequences sequences was ruled out. The correlation was higher in sequences high in G+C content in regions neighboring the region over which MI was calculated. This suggests that the signal at a distance of 6 bp could result from repetitions of G and C nucleotides. Finally, the existence of this increased predictability at a distance of 6 bp did not depend on the precise alignment of the sequences. Instead, statistically, it appears to be a property of human promoter sequences throught the range from -1000 to + 500 bp re: TSS.

Methods

DNA promoter sequences

DNA promoter region sequences for *Homo sapiens* were retrieved using R/Bioconductor software (GenomicFeatures::promoters function) from http://genome.ucsc.edu/ [15]. Data from the February 2009 (GRCh37/hg19) Assembly were accessed on May 1, 2019. For each gene, the bp sequence from -1000 to +600 base pairs (bp) relative to the putative TSS was downloaded. Analyses were performed separately on sequences designated as being from the $+$ *vs.* the $-$ strands. To prevent possible intrusion of harmonics from the exonic 3–bp signal from intruding into the analyses of the promoter sequences, the position of the start codon for each gene used in the two sets of promoter sequences was determined and checked for overlap with the corresponding 1600 bp promoter sequence. Based on this check, 10 bp were removed from the $3'$ end of every promoter sequence to eliminate possible overlap with exonic sequences. This left promoter sequences with lengths of 1590 bp for analysis. After duplicate sequences and sequences containing missing data in the range above were removed, this left 11,602 and 11,177 sequences for the $+$ and $-$ strands, respectively.

CDS sequences were used as a methodological control. Human coding DNA contains a prominent repetitive signal at a distance of 3–bp [3], so the MI measure should be able to detect that signal in CDS sequences. CDS sequences were accessed through Bioconductor and the annotation package TxDb.Hsapiens.UCSC.hg19.knownGene [16]. This produced CDS sequences from a total of 63,691 transcripts (32,291 on the plus strand and 31,400 on the minus strand). The minimum CDS sequences width was 75 bp on both strands. The first 75 bp on both strands of the CDS sequences were used with the MI measure.

Mutual Information

Mutual information (MI) is an information–theoretic measure of the amount of information about a random variable Y that is contained in the random variable X when X and Y are jointly distributed. Its theoretical lower bound is 0: a value that obtains only when X and Y are independent. Unlike a Pearson r correlation coefficient that only reflects the degree of linear relationship between X and Y, MI is sensitive to all relationships including nonlinear ones. It is also especially useful when the random variables X and Y are discrete and their values are not restricted to being numeric, but can also be symbolic (e.g., the DNA letters A, C, G and T).

Consider a collection of N sequences each of length l aligned at their putative Transcription Start Sites (TSSs). With the random variable $X = x \in \{A, C, G, T\}$, a set of such sequences can be represented as

$$\begin{array}{ll} x_{11}\ x_{12} & \dots\ x_{1l} \\ x_{21}\ x_{22} & \dots\ x_{2l} \\ x_{31}\ x_{32} & \dots\ x_{3l} \\ \dots\ \dots & \dots\ \dots \\ x_{N1}\ x_{N2} & \dots\ x_{Nl} \end{array} \tag{1}$$

Here, the first index refers to a particular sequence, and the second index to a position within that sequence. Each row holds a sequence, and each column holds a position within that sequence. To compute MI between a pair of positions across this set of

N sequences, estimates must be obtained of the joint probabilities of the four bases at these two positions as well as the simple probabilities of the four bases at the first position and also at the second position. There are 16 joint probabilities for the four bases at the two positions. If the two positions are designated as w_1 and w_2, then the 16 joint probabilities would be $p(A_{w_1}A_{w_2}),\ p(A_{w_1}C_{w_2}),\ p(A_{w_1}G_{w_2}),\ \ldots p(T_{w_1}G_{w_2}),\ p(T_{w_1}T_{w_2})$. Similarly, the four simple (marginal) frequencies of the four bases at position w_1 would be $p(A_{w_1}),\ p(C_{w_1}),\ p(G_{w_1}),\ p(T_{w_1})$ and similarly for the four base at position w_2.

To estimate these joint and marginal probabilities for positions w_1 and w_2, the $4x4$ joint *frequency* table is first obtained as shown in Table 1. These joint and marginal frequencies are then converted to proportions. For example:

$$\hat{p}(C_{w_1}A_{w_2}) = \frac{f(C_{w_1}A_{w_2})}{N} \tag{2}$$

These proportions are then taken as estimates of the joint and marginal probabilities necessary to estimate MI. This yields an estimated joint probability table like the one shown in Table 2.

Mutual information (MI) is computed between a pair of positions w_1 and w_2 using an estimated joint probability table like the one shown in Table 2. Two positions at a distance of k bp would sit with the first or $5'$ base at position w_1 in the sequence, and the second or $3'$ base at position w_2 giving $k = |w_1 - w_2|$. With the bases at the first of the pair of positions given as $X = x \in \{A,C,G,T\}$ and the bases at the second position in the pair given as $Y = y \in \{A,C,G,T\}$, MI for a distance k is computed as

$$\widehat{MI(k)} = \sum_x \sum_y \hat{p}(x_{w_1}y_{w_2}) log\left(\frac{\hat{p}(x_{w_1}y_{w_2})}{\hat{p}(x_{w_1})\hat{p}(y_{w_1})}\right) \tag{3}$$

The natural log was used in this study. MI then is theoretically bounded by 0 at its low end and by $log(4) = 1.386$. The 4 enters the upper bound equation as the cardinality of the DNA symbol set $\{A,C,G,T\}$.

It should be noted that the finite lengths of the sequences shown in Equation 1 mean that the true joint probability distribution required for the calculation of MI is only approximated by the estimated joint probability distribution. This has two effects: 1) It introduces error variance into the estimate of MI that scales as $\frac{1}{N}$ with N as the number of sequences from which the joint probability distribution was estimated [17], and 2) It introduces an upward bias into this estimate as well [17–19]. For that reason, the bias–corrected measure of MI, *bcMI*, from the R Bioconductor package *mpmi* was used. This measure corrects the bias in MI using a jackknife procedure.

An alternative definition of MI can also prove useful in this analysis. The MI between two random variables, X and Y, can be defined as

$$
\begin{aligned}
& MI(X;Y) = H(X) + H(Y) - H(X;Y) \\
& with \\
& H(X) = -\sum p(X) log(p(X)) \quad entropy\ of\ X \\
& H(Y) = -\sum p(Y) log(p(Y)) \quad entropy\ of\ Y \\
& and \\
& H(X;Y) = -\sum p(X;Y) log(p(X;Y)) \quad joint\ entropy\ of\ X\ and\ Y
\end{aligned}
\tag{4}
$$

The sum is over the four letters of the DNA alphabet.

Variance of the Mutual Information Estimator

The variance of the MI measure was estimated using the method in [17]. Briefly, the variance of MI scales with $1/n$ where n is the number of sequences used in the calculation. The total number of sequences split by strand can be subdivided (not randomly subsampled) into equally–sized subsamples of a range of sizes. For example, there were 11,602 promoter sequences specified on the plus strand. This total can be subdivided into equal size subsamples; it could be subdivided into 2 subsamples each of size 5,801 sequences. However, the only two subsample sizes that are possible are $n = 2$ and $n = 5801$. To estimate the relation between the size of these subsamples and the variance of the MI estimate requires a reasonable number of possible equally–sized subsamples sizes. To achieve this, the total (in this case, 11,602) was allowed to decrease by a maximum of 50 to 11,552 sequences. In the range from 11,552 to 11,602, the number was sought that had the most integer factors. In this case, that number is 11,600. This permits subdivision into 28 different equally–sized subsamples that range in size from from $n = 2\ to\ n = 5800$ sequences. For example, 11,600 can be subdivided into 100 equally–sized subsamples each of size 116 sequences.

The variance of the MI measures is modeled as

$$\sigma^2(N) = \frac{B}{N} \tag{5}$$

where the value of B depends on the properties of the distribution of MI [17]. Using the example above, with a subsample of size $n = 116$, MI is calculated for each of these 100 subsamples. The variance of these 100 MI values then provides one point for estimating a slope and intercept characteristic of B. Because 11,600 can be subdivided into 27 other equally–sized subsamples, there are then 27 additional MI variances that can be used to estimate the slope and intercept that relate the size of the subsample, N in Equation 5, to the variance of the corresponding MI measures. The slope and intercept that characterize B can be estimated by regressing the set of 28 variances against their corresponding subsample sizes [1]. Once this slope and intercept have been been estimated, they are used to extrapolate

[1]The logs of the variances are regressed against the logs of the inverses of the subsample sizes

a variance estimate based on a sample size of 11,600 sequences. Extrapolation is necessary because the largest subsample size, 5,800 sequences, used in the procedure was only half of the total sample size. The minor variation between the true sample size of 11,602 and the operative sample size of 11,600 was judged to be trivial enough to be ignored in terms of its effect on the extrapolated MI variance estimate.

Finally, because mean MI was plotted against distance in base pairs, and because each MI mean was based on 21 different estimates of MI at each distance, the variance of the MI estimate for a given distance described in the preceding paragraph was divided by 21, and the square root of this result was used as an estimate of the standard error of the mean (SEM) in plots such as the one shown in Figure 2.

Mutual Information and G+C content.
To examine the impact of *G+C* content on features of the MI *vs.* distance function, sequences specified on the plus strand were split into two groups: low *vs.* high *G+C* content. To effect this split, the following procedure was used. The proportions of the four nucleotides in the range of positions from -300 to -100 bp re: TSS was computed separately for each of the 11,602 sequences. These proportions were then modeled as two–component mixtures of multinomial distributions using the *multmixEM* function in the R package *mixtools*. The algorithm estimates the proportions of the four bases in these two components across the entire sample of sequences. Component 1 had slightly higher A+T than G+C content. Component 2 had higher G+C content than A+T content. The posterior probability of each of the two components for each sequence is then used to classify each sequence as high *vs.* low G+C content.

$$\begin{aligned} \lambda < 0.2 &\Rightarrow low\ \ G+C \qquad n = 4974\ \ sequences \\ \lambda < 0.8 &\Rightarrow high\ \ G+C \qquad n = 6065\ \ sequences \end{aligned} \tag{6}$$

MI was then calculated for distances from 1 to 21 bp relative to an initial position of -99 bp re: TSS. Note, this range over which the *G+C* content was tallied is just $5'$ of the range of positions used to estimate the MI *vs.* distance function. In this way, the statistics of the *G+C* content and of MI *vs.* distance did not depend directly on ranges of positions that overlapped. The plots of MI *vs.* distance were then examined separately for the low and high *G+C* content sequences to determine the possible impact of this more global parameter on the presence of the increase in MI at a distance of 6 bp.

Declarations
Ethics approval and consent to participate
Not applicable.

Availability of data and material
The promoter sequences analyzed during the current study were contained in the TxDb.Hsapiens.UCSC.hg19.knownGene repository and were retrieved using R/Bioconductor *GenomicFeatures* from https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19.

Competing interests
The author declares that he has no competing interests.

Funding
This research was supported by funds provided to the author by the Lynette S. Autrey Foundation, Rice University. The Lynette S. Autrey Foundation had no role in in the design of the study, in the collection, analysis, and interpretation of data, nor in writing the manuscript.

Authors' contributions
Not applicable.

Acknowledgements
I thank Robert L. Urqhart for additional computational support and for discussions related to mutual information.

Author's information
The author holds a Lynette S. Autrey Chair at Rice University. Portions of this work were presented at the annual conference of the American Society of Human Genetics, October, 2019, Houston, TX, USA.

References

1. Datta, S., Asif, A.: A fast dft based gene prediction algorithm for identification of protein coding regions. In: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 5, p. 653 (2005). IEEE
2. Grosse, I., Herzel, H., Buldyrev, S.V., Stanley, H.E.: Species independence of mutual information in coding and noncoding dna. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics **61**(5 Pt B), 5624–9 (2000)
3. Yin, C., Yau, S.S.-T.: Prediction of protein coding regions by the 3-base periodicity analysis of a dna sequence. J Theor Biol **247**(4), 687–94 (2007). doi:10.1016/j.jtbi.2007.03.038
4. Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L.: Study of statistical correlations in dna sequences. Gene **300**(1-2), 105–15 (2002)
5. Holste, D., Grosse, I., Herzel, H.: Statistical analysis of the dna sequence of human chromosome 22. Phys Rev E Stat Nonlin Soft Matter Phys **64**(4 Pt 1), 041917 (2001). doi:10.1103/PhysRevE.64.041917
6. Berryman, M.J., Allison, A., Abbott, D.: Mutual information for examining correlations in dna. Fluctuation and Noise Letters **4**(02), 237–246 (2004)
7. Li, W., Kaneko, K.: Long-range correlation and partial $1/f$ spectrum in a noncoding dna sequence. EPL (Europhysics Letters) **17**(7), 655 (1992)
8. Gymrek, M.: A genomic view of short tandem repeats. Curr Opin Genet Dev **44**, 9–16 (2017). doi:10.1016/j.gde.2017.01.012
9. Herzel, Ebeling, Schmitt: Entropies of biosequences: The role of repeats. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics **50**(6), 5061–5071 (1994)
10. Madsen, B.E., Villesen, P., Wiuf, C.: A periodic pattern of snps in the human genome. Genome Res **17**(10), 1414–9 (2007). doi:10.1101/gr.6223207
11. Klug, A., Lutter, L.C.: The helical periodicity of dna on the nucleosome. Nucleic Acids Res **9**(17), 4267–83 (1981). doi:10.1093/nar/9.17.4267
12. Levitt, M.: How many base-pairs per turn does dna have in solution and in chromatin? some theoretical calculations. Proc Natl Acad Sci U S A **75**(2), 640–4 (1978)
13. Gabrielian, A., Pongor, S.: Correlation of intrinsic dna curvature with dna property periodicity. FEBS Lett **393**(1), 65–8 (1996)
14. Sueoka, N.: Intrastrand parity rules of dna base composition and usage biases of synonymous codons. J Mol Evol **40**(3), 318–25 (1995). doi:10.1007/bf00163236
15. Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G.P., Haussler, D., Kuhn, R.M., Kent, W.J.: The ucsc genome browser database: 2019 update. Nucleic Acids Res **47**(D1), 853–858 (2019). doi:10.1093/nar/gky1095
16. Carlson, M., Maintainer, B.: Txdb. hsapiens. ucsc. hg19. knowngene: Annotation package for txdb object (s). In: R Package Version 3.2.2, (2015)
17. Holmes, C.M., Nemenman, I.: Estimation of mutual information for real-valued data with error bars and controlled bias (2019)
18. Zhu, J., Bellanger, J.-J., Shu, H., Yang, C., Le Bouquin Jeannès, R.: Bias reduction in the estimation of mutual information. Phys Rev E Stat Nonlin Soft Matter Phys **90**(5-1), 052714 (2014). doi:10.1103/PhysRevE.90.052714
19. Zhang, Z., Zheng, L.: A mutual information estimator with exponentially decaying bias. Stat Appl Genet Mol Biol **14**(3), 243–52 (2015). doi:10.1515/sagmb-2014-0047

1 Figures

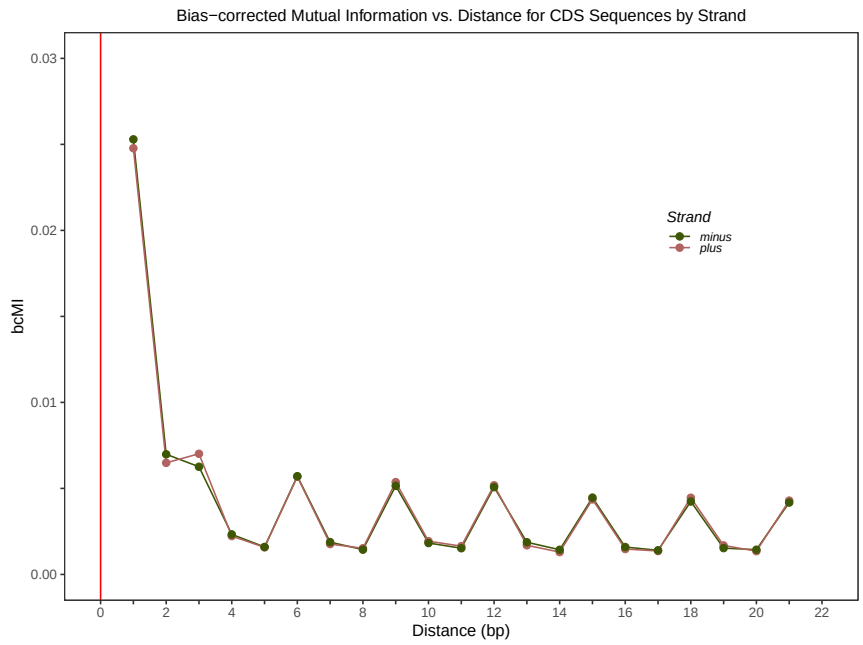

Figure 1 bcMI vs. Distance: CDS Bias-corrected mutual information for CDS sequences by strand. The red vertical line is the relative position of the initial base from which MI at all distances was computed.

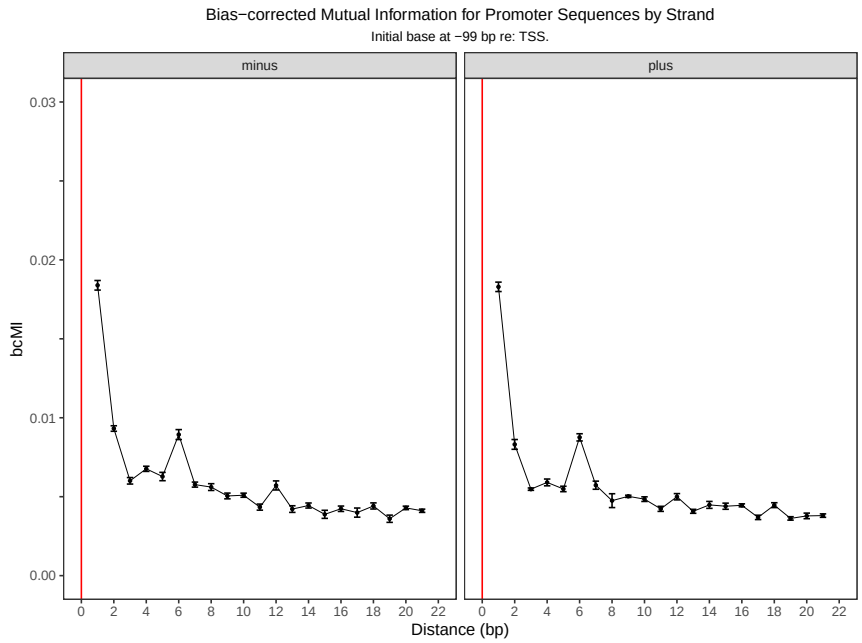

Figure 2 bcMI vs. Distance: Plus and Minus Strands Bias-corrected mutual information for promoter sequences by strand. The red vertical line is the relative position of the initial base. Error bars are ± 2 SEM.

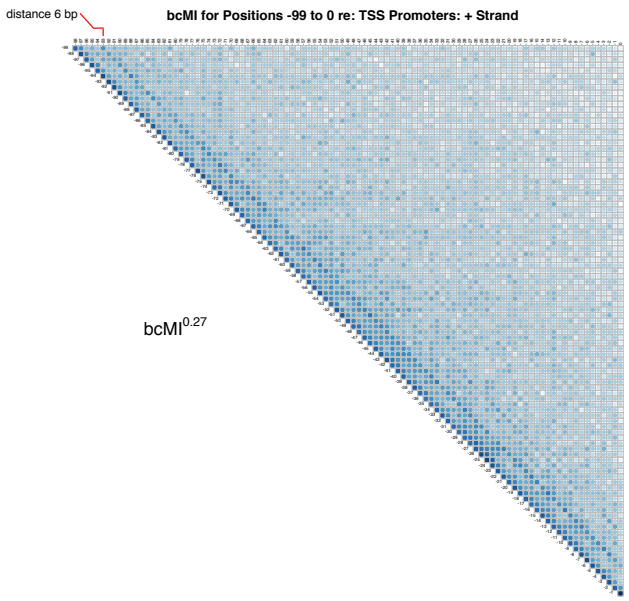

Figure 3 bcMI Correlogram Upper triangle of the matrix holding bcMI for all pairwise positions between -99 and 0 bp re: TSS , + strand. The saturation of the color is proportional to $bcMI^{0.27}$. The bcMI values have been raised to the 0.27 power (a monotonic transformation) to better visualize the signal at a distance of 6 bp. Without this transformation, the figure is dominated by the much stronger signal at 1 bp.

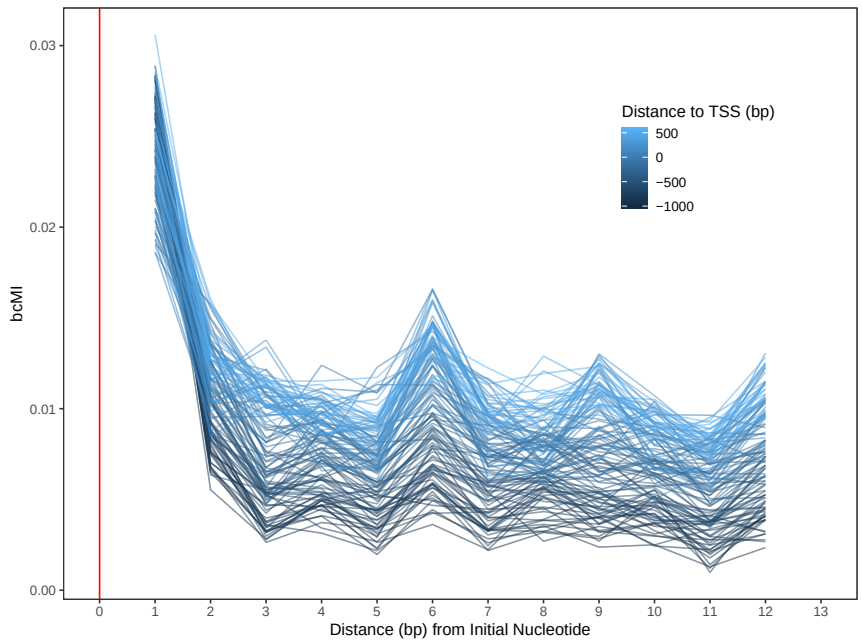

Figure 4 bcMI vs. distance by starting position bcMI vs. Distance to second nucleotide for plus strand sequences. Each curve represents a different distance between the initial nucleotide (0 on the x–axis) and the TSS with the lowest curve having an initial nucleotide at -1000 bp from the TSS. For each subsequent curve the starting position shifts 13 bp toward the TSS and the vertical position of the curve has been shifted by an additional 0.00005 MI units. The initial nucleotide from which all MI values were calculated was always positioned at 0 (red line) relative to the distance to subsequent positions.

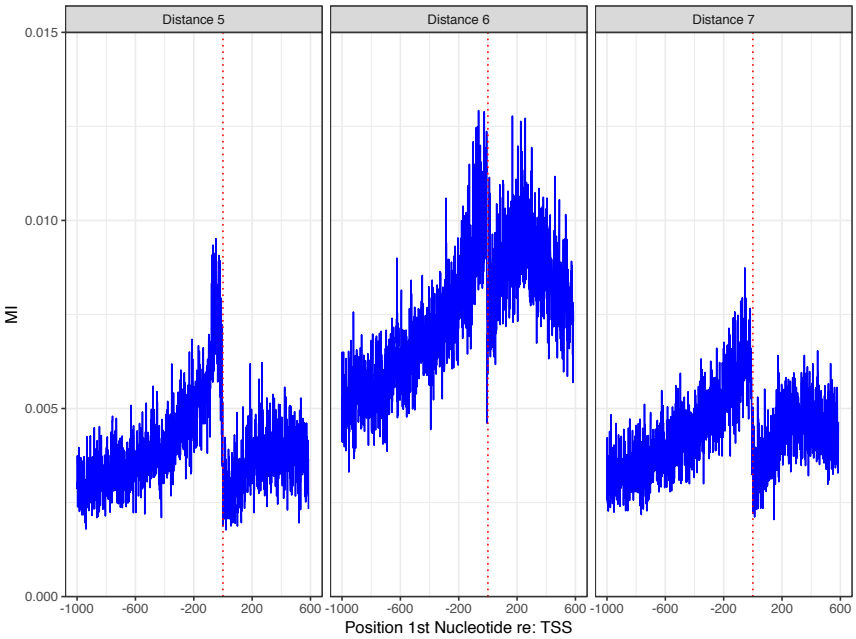

Figure 5 bcMI at distances 5, 6 and 7, base pairs bcMI is shown for distances 5, 6 and 7 base pairs for plus strand sequences as the initial nucleotide steps from 1000 bp upstream of the TSS to 574 base pairs downstream of the TSS.

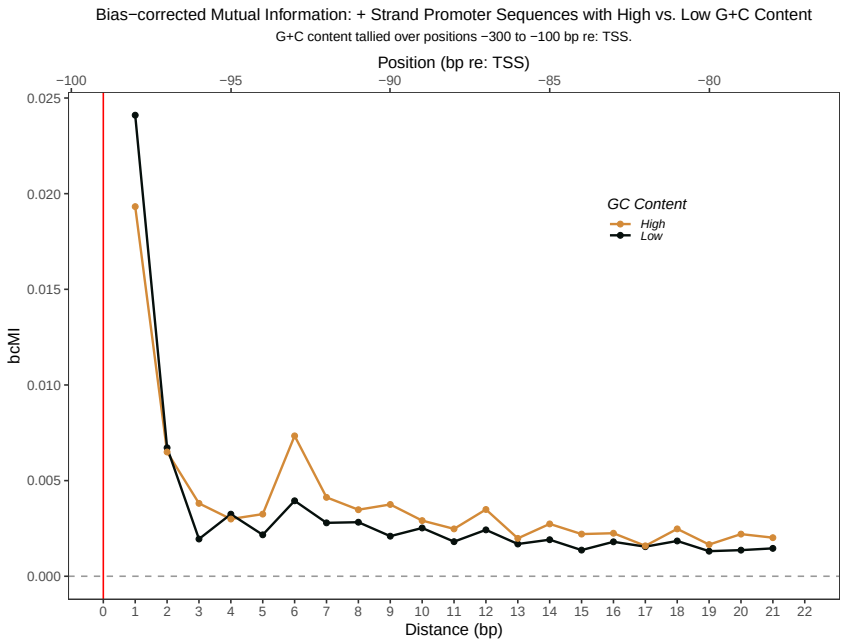

Figure 6 bcMI split by GC content bcMI is shown for sequences coded on the plus strand split by G+C content. G+C content was tallied over the range to -300 to -100 bp re: TSS. There are 6192 and 5410 sequences with low G+C *vs.* high G+C content, respectively.

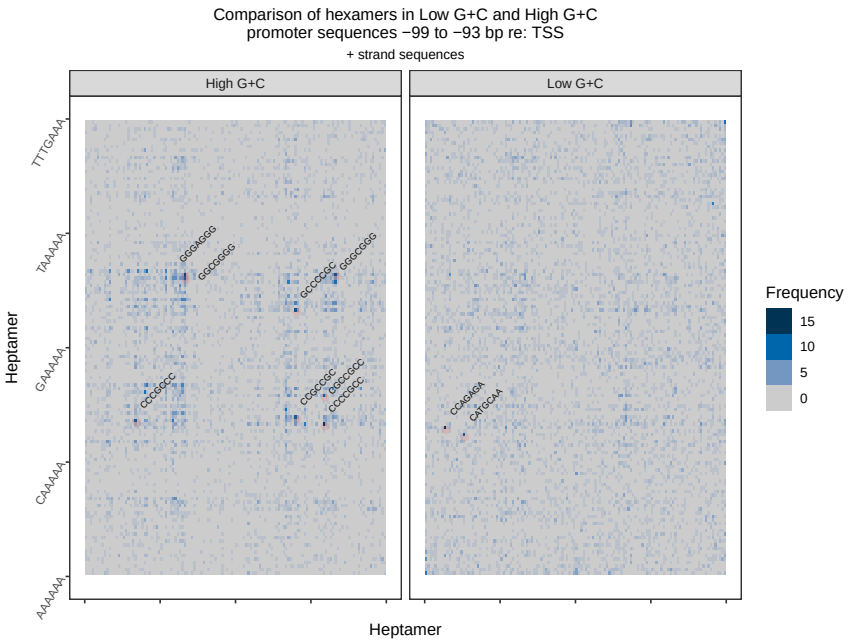

Figure 7 Most frequent heptamers Heptamer frequencies in high *vs.* low G+C content promoters (plus strand only) in the region from -300 to -100 bp re: TSS. All possible heptamers are indicated by their positions in a 128 x 128 grid. At the lower left in each panel is the heptamer AAAAAAA. Progressing toward the right the rightmost base cycles the fastest through the four bases, followed by the base second from the right etc. The heptamer at the lower right corner in each panel is the 128th in sequence: AAACTTT. The set continues in the second row from the bottom left until all 16,384 heptamers have been located in the grid. The frequency with which each hexamer appeared across the 11,602 sequences in the range from -300 to -100 re: TSS is indicated by the saturation of the point. The frequencies based on sequences classified with high G+C content are shown on the left, and based on those with low G+C content are indicated on the right. Heptamers that appeared with a frequency greater than or equal to 10 tallied across all 11,602 sequences are highlighted in red, and their 7-base sequences label these points. Seven of the eight more frequent heptamers show the same base in the first and seventh positions (distance = 6 bp) in the sequences with high G+C content, while neither of the more frequent heptamers from the low G+C content sequences show a repeat of the first base in the seventh position.

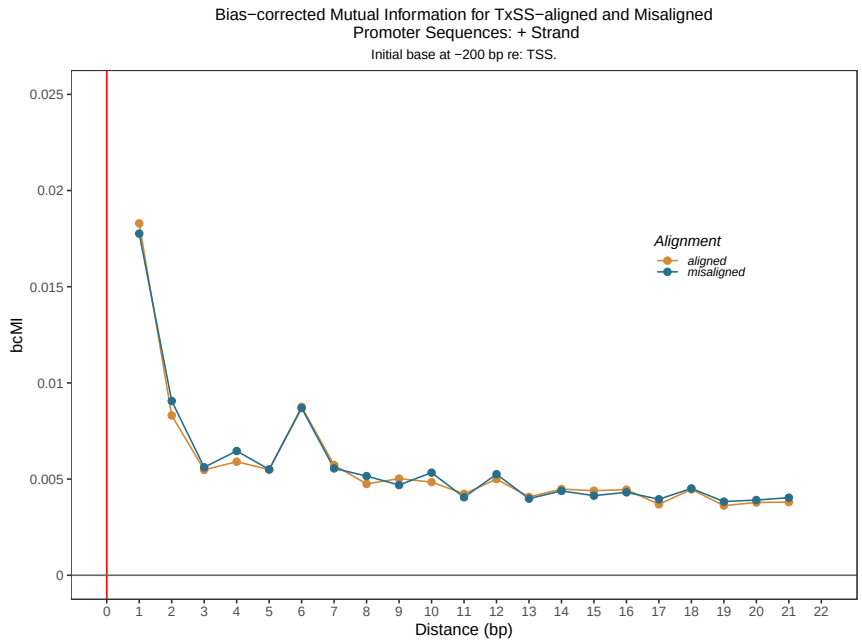

Figure 8 bcMI aligned *vs.* misaligned sequences bcMI is shown for the original sequences (plus strand) aligned at their putative TSS and for those sequences purposely misaligned by a random integer step in the range from ± 40 bp. The misalignment has essentially no effect on the curve relating bcMI to distance.

Tables

Table 1 Joint frequencies of the four DNA bases at two locations w_1 *and* w_2 across N aligned sequences.

		w_2				
		A	C	G	T	$f(\cdot w_1)$
w_1	A	$f(A_{w_1}A_{w_2})$	$f(A_{w_1}C_{w_2})$	$f(A_{w_1}G_{w_2})$	$f(A_{w_1}T_{w_2})$	$f(A_{w_1})$
	C	$f(C_{w_1}A_{w_2})$	$f(C_{w_1}C_{w_2})$	$f(C_{w_1}G_{w_2})$	$f(C_{w_1}T_{w_2})$	$f(C_{w_1})$
	G	$f(G_{w_1}A_{w_2})$	$f(G_{w_1}C_{w_2})$	$f(G_{w_1}G_{w_2})$	$f(G_{w_1}T_{w_2})$	$f(G_{w_1})$
	T	$f(T_{w_1}A_{w_2})$	$f(T_{w_1}C_{w_2})$	$f(T_{w_1}G_{w_2})$	$f(T_{w_1}T_{w_2})$	$f(T_{w_1})$
	$f(\cdot w_1)$	$f(A_{w_{12}})$	$f(C_{w_2})$	$f(G_{w_2})$	$f(T_{w_2})$	N

Table 2 Estimated joint and marginal probabilities ($\hat{p}$) of the four DNA bases at two locations w_1 *and* w_2 across N aligned sequences.

		w_2				
		A	C	G	T	$f(\cdot w_1)$
w_1	A	$\hat{p}(A_{w_1}A_{w_2})$	$\hat{p}(A_{w_1}C_{w_2})$	$\hat{p}(A_{w_1}G_{w_2})$	$\hat{p}(A_{w_1}T_{w_2})$	$\hat{p}(A_{w_1})$
	C	$\hat{p}(C_{w_1}A_{w_2})$	$\hat{p}(C_{w_1}C_{w_2})$	$\hat{p}(C_{w_1}G_{w_2})$	$\hat{p}(C_{w_1}T_{w_2})$	$\hat{p}(C_{w_1})$
	G	$\hat{p}(G_{w_1}A_{w_2})$	$\hat{p}(G_{w_1}C_{w_2})$	$\hat{p}(G_{w_1}G_{w_2})$	$\hat{p}(G_{w_1}T_{w_2})$	$\hat{p}(G_{w_1})$
	T	$\hat{p}(T_{w_1}A_{w_2})$	$\hat{p}(T_{w_1}C_{w_2})$	$\hat{p}(T_{w_1}G_{w_2})$	$\hat{p}(T_{w_1}T_{w_2})$	$\hat{p}(T_{w_1})$
	$\hat{p}(\cdot w_1)$	$\hat{p}(A_{w_{12}})$	$\hat{p}(C_{w_2})$	$\hat{p}(G_{w_2})$	$\hat{p}(T_{w_2})$	N