A model of *k*-mer surprisal to quantify local sequence information content surrounding splice regions

Sam Humphrey[1,2], Alastair Kerr[1,2], Magnus Rattray[3], Caroline Dive[1,2] and Crispin J. Miller[4,5]

[1] CRUK Manchester Institute Cancer Biomarker Centre, The University of Manchester, Manchester, United Kingdom
[2] CRUK Manchester Institute, CRUK Lung Cancer Centre of Excellence, Manchester, United Kingdom
[3] Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, United Kingdom
[4] Computational Biology Group, CRUK Beatson Institute, Glasgow, United Kingdom
[5] Institute of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom

ABSTRACT

Molecular sequences carry information. Analysis of sequence conservation between homologous loci is a proven approach with which to explore the information content of molecular sequences. This is often done using multiple sequence alignments to support comparisons between homologous loci. These methods therefore rely on sufficient underlying sequence similarity with which to construct a representative alignment. Here we describe a method using a formal metric of information, surprisal, to analyse biological sub-sequences without alignment constraints. We applied our model to the genomes of five different species to reveal similar patterns across a panel of eukaryotes. As the surprisal of a sub-sequence is inversely proportional to its occurrence within the genome, the optimal size of the sub-sequences was selected for each species under consideration. With the model optimized, we found a strong correlation between surprisal and CG dinucleotide usage. The utility of our model was tested by examining the sequences of genes known to undergo splicing. We demonstrate that our model can identify biological features of interest such as known donor and acceptor sites. Analysis across all annotated coding exon junctions in *Homo sapiens* reveals the information content of coding exons to be greater than the surrounding intron regions, a consequence of increased suppression of the CG dinucleotide in intronic space. Sequences within coding regions proximal to exon junctions exhibited novel patterns within DNA and coding mRNA that are not a function of the encoded amino acid sequence. Our findings are consistent with the presence of secondary information encoding features such as DNA and RNA binding sites, multiplexed through the coding sequence and independent of the information required to define the corresponding amino-acid sequence. We conclude that surprisal provides a complementary methodology with which to locate regions of interest in the genome, particularly in situations that lack an appropriate multiple sequence alignment.

Subjects Computational Biology, Genomics, Mathematical Biology, Computational Science
Keywords Information theory, Surprisal, Splicing, Entropy

Submitted 27 April 2020
Accepted 8 September 2020
Published 4 November 2020

Corresponding authors
Sam Humphrey,
Sam.Humphrey@postgrad.manchester.ac.uk
Crispin J. Miller,
crispin.miller@glasgow.ac.uk

Academic editor
Thomas Tullius

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.10063**

How to cite this article Humphrey S, Kerr A, Rattray M, Dive C, Miller CJ. 2020. A model of *k*-mer surprisal to quantify local sequence information content surrounding splice regions. *PeerJ* **8**:e10063 http://doi.org/10.7717/peerj.10063

RATIONALE

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection (*Dayhoff, Schwartz & Orcutt, 1978*; *Dayhoff, Barker & Hunt, 1983*). Genomic regions conserved between species therefore constitute islands of evolutionary stability within the more rapidly evolving nucleic acid sequence, and thus represent loci where important features are encoded. These observations make it possible to study evolutionary processes by generating multiple sequence alignments that seek to characterise the genetic changes that occur over evolutionary timescales. Methods such as the MEME and DREME suite of tools (*Bailey & Elkan, 1995*; *Bailey, 2011*), identify significant encodings using statistically enriched motifs in sets of functionally related molecular sequences. Here we have developed a complementary method to identify important sequence encodings within molecular sequences. Our approach measures the information provided by sub-sequences surrounding individual loci and we have shown this method can identify important genomic features. This method is alignment free, and hence can be applied broadly across all sequences, irrespective of overall sequence similarity, and independent of the functional relationships that might be used to group them. The approach can also evaluate different types of molecular sequences such as coding sequences and amino acids. Further applications of this approach therefore include the analysis of seemingly unconnected genomic loci such as those harbouring single nucleotide variants (SNVs) or somatic mutations.

INTRODUCTION

Information theory

In 1948, Shannon linked the information content of a sequence of symbols, first described by Hartley, and entropy, a quantity used in thermodynamics (*Shannon, 1948*; *Hartley, 1928*; *Gibbs, 1902*). Shannon's discoveries initially focussed on transmission of messages over noisy channels. These later became fundamental principles in information storage (*MacKay, 2003*). Shannon Entropy is a measure of the complexity of an ensemble X, of symbols x, where each symbol occurs with a probability $p(x)$. The self-information associated with each symbol is called *surprisal* (*Tribus, 1961*) and is defined by:

$$S(x) = -\log_2(p(x)) \tag{1}$$

where $S(x)$ is measured in bits. For the full ensemble of n symbols, the total information of the ensemble is the sum over all surprisals $I(X) = \sum_{i=1}^{n} S(x_i)$. The Shannon entropy is defined as the average information per symbol or the expectation value of all surprisals:

$$H(X) = E(S(x)) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i)) \tag{2}$$

where $\sum_{i=1}^{n} p(x_i) = 1$ and $H(X)$ is also measured in bits. A few years after this formulation, the structure of DNA and the first protein sequences were discovered (*Sanger, 1952*; *Watson & Crick, 1953*). These discoveries and further advances in biology enabled the application of information theory to biological sequences. The storage of biological information within

ensembles of molecular sequences led to several investigations of the quantification of biological information and its association with with biological functions (*Gatlin, 1966*). The Central Dogma of molecular biology itself was first framed in terms of the "transfer of sequential information" (*Crick, 1970*; *Cobb, 2017*), and information theory continues to underpin our understanding of how the genome encodes genetic information (*Pritišanac et al., 2019*).

A DNA sequence can be represented using an alphabet corresponding to the individual nucleotides $\mathbb{A} = \{A, C, G, T\}$. Since each amino acid is defined by a tri-nucleotide, or codon, it is useful to consider nucleotide sequences using an alphabet of 64 symbols, one for every possible tri-nucleotide. Of these, 61 codons encode amino acids, and the remaining three correspond to stop codons. Similarly, protein sequences can be defined by an alphabet of 20 symbols, one for each amino acid. This decline in the number of possible symbols between coding DNA, RNA, and protein leads to a decline in the maximum amount of information that can be encoded at each level. Information theory provides a theoretical framework within which to quantify these differences (*Yockey, 1974*; *Nemzer, 2017*).

The reason why almost all organisms translate only 20 amino acids despite the ability to encode 61 possible codons has not been fully determined (*Koonin & Novozhilov, 2009*). However, the redundancy in the genetic code allows for additional information to be captured within a coding region beyond that required to define the amino acid sequence itself (*Yockey, 2000*; *Itzkovitz, Hodis & Segal, 2010*). It has been suggested that codon degeneracy and the structure of the codon table support functions in addition to the encoding of amino acids (*Dayhoff, Schwartz & Orcutt, 1978*; *Henikoff & Henikoff, 1992*; *Itzkovitz & Alon, 2007*; *Berleant et al., 2009*; *Maraia & Iben, 2014*). These include the description of splicing regulatory motifs (*Lim & Burge, 2001*; *Zhang & Chasin, 2004*; *Wang & Burge, 2008*), DNA binding sites that co-exist within the coding sequence (*Melnik & Usatenko, 2014*; *Vinga, 2014*; *Shreif, Striegel & Periwal, 2015*), and RNA secondary structure (*Itzkovitz, Hodis & Segal, 2010*). This additional information can be viewed as a separate signal containing non-coding information multiplexed through the protein coding sequence.

While the genetic code naturally leads to a focus on triplet sequences, other representations are possible, and different length sequences reveal different aspects of the genome. For example, the information associated with individual nucleotides can be used to identify the presence of motifs within an ensemble of short molecular sequences (*Schneider et al., 1986*). It is important to consider that information content shown in motif figures is usually represented as $2 - H(X)$ as the aim is to identify consistent nucleotides within the motif rather than diversity (*Schneider & Stephens, 1990*). Dinucleotides can also be used to represent DNA sequences as they are important in the specification of epigenetic modifications (CpG islands), binding sites, splice donors and acceptors. Alphabets representing molecular sub-sequences of length k (k-mers) have also been widely used in motif discovery (*Castle et al., 2008*; *Bailey, 2011*). Information theoretic approaches have previously been used to investigate biological features such as coding and non-coding regions, nucleosome positioning and DNA binding sites with a

variety of methods and representations to quantify information (*Koslicki, 2011*; *Vinga, 2014*; *Wu, Zhang & Mu, 2014*).

Splicing

Splicing is an essential mechanism in human cells performed by the spliceosome, a large ribonucleoprotein complex comprised of five small ribonucleoproteins at its core plus many other protein cofactors (*Matera & Wang, 2014*). The process of splicing involves precise removal of intron sequences from the transcribed RNA sequence. Each gene can express multiple mRNAs with different patterns of intron removal, such that exons in one mRNA may be part of an intron for another. Approximately 95% of genes are spliced in humans through this exquisitely regulated process, which is responsible for much of the diversity in the proteome (*Lee & Rio, 2015*; *Sahebi et al., 2016*). For a mechanistic review see *Shi (2017)*. Exon splicing is determined by the binding of three key molecules: U1, splicing factor 1 (SF1) and the U2 auxiliary factor (U2AF) to the 5′ splice site, the branch point and 3′ splice site within the intron, respectively. The decision to include an exon within a transcript is generally made at the time of binding of these three molecules and is mediated by the use of Splicing Regulatory Elements (SREs) (*Sickmier et al., 2006*; *Diederichs et al., 2016*; *Saha et al., 2020*). Some of these SREs are embedded within exons and can be viewed as loci where additional information is multiplexed along with the information required for amino acid sequence determination. Analysis of recurrent k-mers near splice junctions have already been shown to predict novel splice sites and sequences involved in splicing (*Lim & Burge, 2001*; *Zhang & Chasin, 2004*; *Fairbrother et al., 2004*; *Schwartz, Hall & Ast, 2009*; *Raponi et al., 2011*; *Ke et al., 2011*; *Erkelenz et al., 2014*).

Here we have created a method whereby different biological sequences can be interrogated without requiring a direct multiple sequence alignment, or the need for sufficient sequence conservation with which to build that alignment. Using splice sites as a set of known biological features, we show that our model can quantify the information content of sequences at these regions irrespective of species.

METHODS

Annotation

Genomic annotation for 5 different species *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Danio rerio*, *Schizosaccharomyces pombe* was downloaded from Ensembl v99 (ftp://ftp.ensembl.org/pub/) (*Cunningham et al., 2019*). Genomic DNA, coding mRNA, and peptide files, were mapped against genomic annotation provided by the gene transfer format (.gtf) file. Only protein coding transcripts with GENCODE basic annotation were included in this analysis, which is defined as the set of 5′ and 3′ complete transcripts. Exon-intron boundaries and exon-exon junctions were independently generated using the gene transfer format annotation and sequences were aligned at splice sites.

Only k-mers with nucleotide symbols consisting fully of $\mathbb{A}_{nt} = \{A, C, G, T\}$ or $\mathbb{A}_{aa} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ for amino acids were used in this analysis. Since the majority of eukaryote genes have multiple transcripts per gene, with many exons shared between them, when evaluating coding sequences and amino acids,

the k-mers from those shared exons would be duplicated if all transcripts were considered individually. We therefore impose a multiple transcript correction such that the genomic locations; chromosome, strand, start and end positions, for all k-mers must be unique. The repeated k-mers used by different transcripts at the same genomic locus are discarded. This method is also sensitive to k-mers spanning exon-exon junctions, since it maps the positions to genomic locations. The coding mRNA distribution contains all k-mers from within coding sequences only; introns, UTRs, and non-coding transcripts were excluded. Amino acid k-mers were reduced in length by a factor of 3, ($k_{aa} = \frac{k_{nt}}{3}$), and their value was allocated to all three nucleotides in their corresponding codon when comparing with DNA and coding mRNA sequences.

Probability of *k*-mer occurrences

The k-mers starting at every position in the DNA (both strands), all coding mRNAs, and all amino acid sequences were extracted, counted and recorded in frequency tables. These frequency tables were then used to identify the probability of the k-mer, x, occurring in the overall sequence

$$p(x) = \frac{counts(x)}{\sum_x counts(x)} \tag{3}$$

where $\sum_x counts(x)$ is equivalent to the total number of k-mers in the sequence, which corresponds to the sequence length minus k and any discarded sequences. Using this measure of k-mer probability, the surprisal for each k-mer can be calculated by Eq. (1) while the Shannon entropy for the total sequence can be found using Eq. (2).

Choice of *k*-mer length

This estimate of k-mer probabilities (Eq.(3)) works well for small k, however for larger k this entropy calculation is limited due to the finite sample size of biological sequences (*Herzel, Ebeling & Schmitt, 1994*; *Herzel & Große, 1995*). Since the total number of possible k-mers in nucleotide space is $4^{k_{nt}}$ and is $20^{k_{aa}}$ in amino acid space, the *H. sapiens* DNA sequence is too small to contain every possible k-mer for $k_{nt} \geq 17$ in DNA sequences and $k_{aa} \geq 6$ for amino-acids. The k-mer distributions associated with these spaces are also skewed, such that some k-mers do not occur even for much smaller k. Here we refer to k-mers with zero occurrences as 'nullomers' (*Hampikian & Andersen, 2007*). The presence of these nullomers at short k, combined with the skew in the distribution, lowers the observed entropy H^{obs} of the sequence. This information loss can be quantified in terms of the sequence redundancy:

$$R = 1 - \frac{H^{obs}}{H^{max}} \tag{4}$$

where the largest possible entropy of the system, $H_{nt}^{max} = -\log_2(4^{-k_{nt}}) = 2k_{nt}$ bits for nucleotide sequences and $H_{aa}^{max} = -\log_2(20^{-k_{aa}})$ bits for amino acid sequences. Since H^{max} is dependent only on k whilst H^{obs} is limited by the length of the sequence, R increases as the total number of possible k-mers becomes greater than sequence size. This effect is also dependent on the uniformity of the k-mer distribution since H is maximised when all k-mers are equally likely.

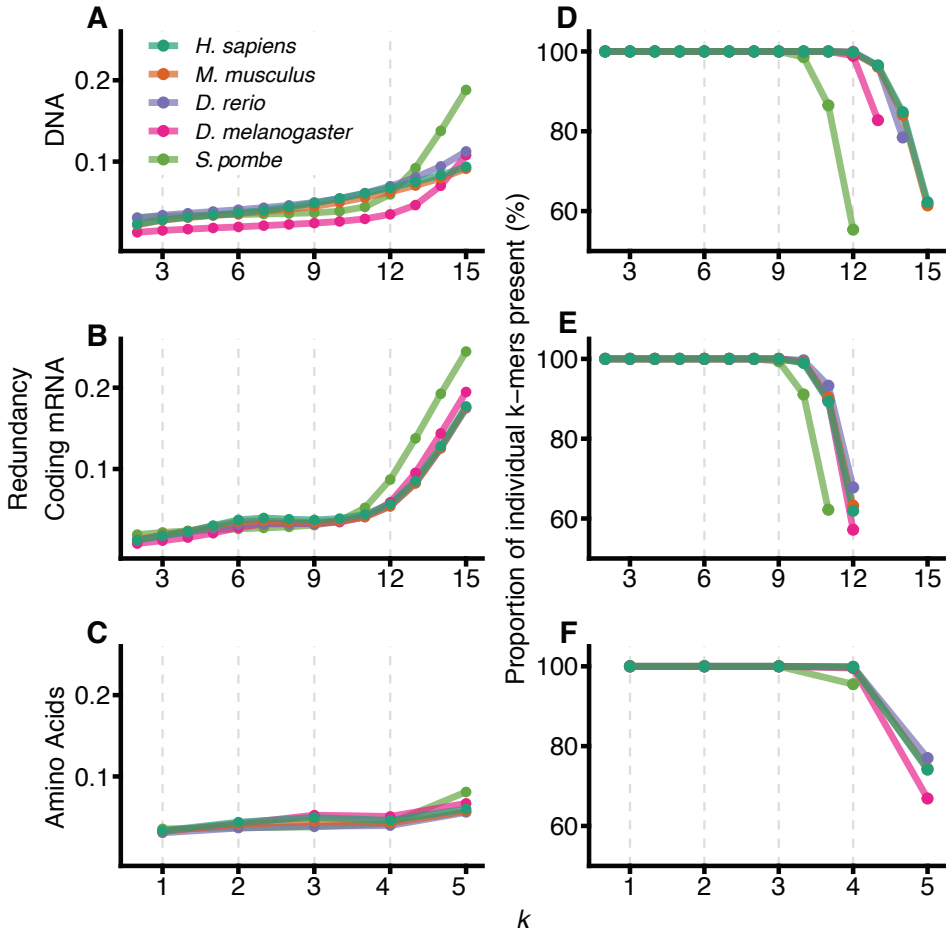

Figure 1 **Visualising the sample size effect.** The effect of increasing k-mer resolution for DNA (A, D), coding mRNA (B, E) and amino acid (C, F) sequences, where amino acids are considered at sizes $\frac{k}{3}$. For each value of k, $2 \leq k \leq 15$, the redundancy (Eq. (4)) of the region was calculated (A, B, C) along with the percentage of all possible k-mers observed at least once in the genome (D, E, F). D, E and F have been truncated at the point where the total possible number of k-mers, $4^{k_{nt}}$ or $20^{k_{aa}}$ exceeds the size of the region for that genome ($\sum_x counts(x)$ for all k-mers, x).

Full-size DOI: 10.7717/peerj.10063/fig-1

Figure 1 shows the redundancy and the proportion of unique k-mers observed for $2 \leq k \leq 15$ for the 5 different species: *H. sapiens*, *M. musculus*, *D. rerio*, *D. melanogaster* and *S. pombe*. As expected, all species show an increase in redundancy at larger k and a corresponding decrease in the number of k-mers represented at least once in the sequence. In DNA space (Figs. 1A and 1D), this effect becomes prominent at $k_{nt} > 12$. We therefore conclude that 12-mers are an appropriate size to model DNA sequences for all species except *S. pombe*, where they are too large relative to the size of the significantly smaller genome. We have therefore removed *S. pombe* from further analysis. Following similar reasoning, for comparisons between DNA, coding mRNA and amino acid sequences, we selected $k_{nt} = 9$ and $k_{aa} = 3$ to reflect the reduced amount of coding mRNA sequence relative to the DNA.

RESULTS

Surprisal patterns across splice sites

Metazoan exon-intron junctions have known conserved features within introns. Downstream of the 3′ end of the vast majority of exons is the GT donor, and upstream of the 5′ end of the exon is a corresponding AG acceptor site. A pyrimidine rich region occurs approximately 4–20 nt upstream of the acceptor site and a conserved but location variable branch point occurs approximately 15–55 nt upstream of the acceptor site (*Corvelo et al., 2010*). These motifs make exon-intron junctions ideal candidates for testing our model. Figure 2 shows the mean of DNA 12-mer surprisal across all protein coding exon-intron junctions for each species. A consistent surprisal pattern is observed, with substantial changes in the mean surprisal at the position of the polypyrimidine tract and splice site motifs. This result is expected since the surprisal, and hence information, is inversely proportional to k-mer occurrence, and therefore surprisal decreases in the presence of common sequences motifs. Conversely, Fig. 2 also reveals that for all species, exons contain significantly greater information than introns near splice boundaries. Although apparently intuitive, it is surprising since exon sequences are constrained to contain only those sequence patterns capable of representing a functional protein, while intron sequences are under no such constraint. For example, k-mers featuring an in-frame stop codon are not permissible within a coding sequence. A naïve view of coding space, therefore, is one in which exon sequences are constructed from a subset of possible k-mers while intron sequences can be constructed from the entire repertoire of k-mers. This predicts that exons would be constructed from more common sequences, and that their surprisal would therefore be lower than that of the corresponding intronic space. By contrast, these data suggest a greater sequence constraint on intronic regions near splice junctions than similar proximal exon regions. This in turn implies that there are additional constraints for intronic sequences, such as those arising from the need to encode intronic SREs. It is tempting to speculate that these patterns result from a selection pressure that excludes certain valid coding sequences from intronic space. Importantly these patterns are robust against different values of k, demonstrating the generality of the model (Fig. S1). Here we show data for $k = 12$, since these have a lower coefficient of variation than those of smaller k (Fig. S2).

Surprisal across exon junctions

Since the distinct patterns in Fig. 2 arise from differences between coding and non-coding sequence spaces, we aimed to identify additional information encoded in the DNA and coding sequences beyond that which is required simply to define the encoded protein sequence. Protein coding exon-exon junctions were aligned at 9-mer resolution (amino acid 3-mers), with $k = 9$ chosen to account for the smaller sample size of coding sequences (Fig. 3). The mean surprisal for *H. sapiens* surrounding exon-exon junctions is 18.0 bits in DNA and 17.4 bits for coding mRNA, which as expected, is slightly lower due to the loss of in-frame stop codon sequences. Amino acid sequences show a significant reduction of approximately 1/3 information when compared to DNA sequences (12.6 bits). This is in keeping with previous work that considers the entropy of codons (*Yockey, 1974*). All

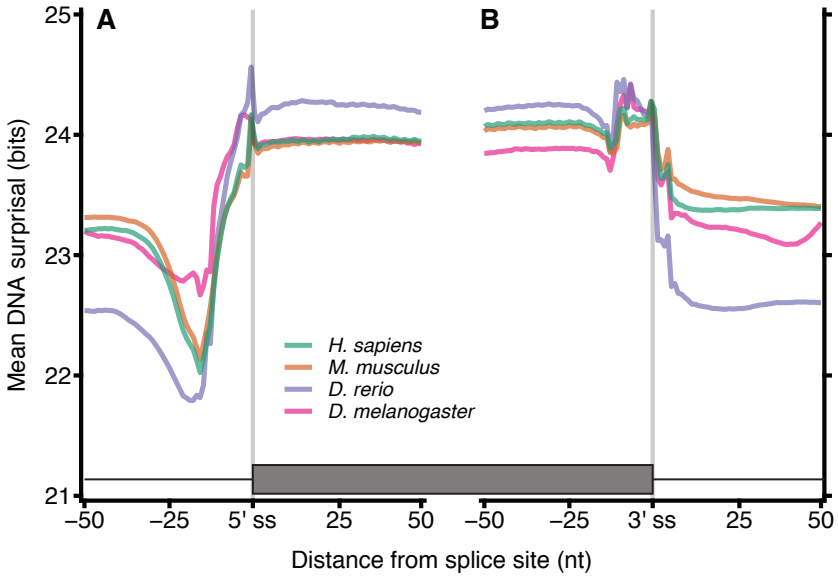

Figure 2 DNA Surprisal across exon-intron junctions identifies known features and reveals increased information within exons. For each of the 4 species, 12-mers were extracted at each position relative to the exon-intron junction for all protein coding exons. The mean surprisal of the 12-mers is plotted across the exon 5′ (A) and 3′ (B) splice site boundaries. Data were generated for all GENCODE basic, protein coding transcripts, using exon annotation downloaded from Ensembl (*Cunningham et al., 2019*). In both plots, the exon-intron junction occurs between positions −1 and 0, indicated by the grey line at position −0.5. Exon junctions falling within 100 nt of the transcript start or end site were removed.

Full-size DOI: 10.7717/peerj.10063/fig-2

other species considered here are consistent with *H. sapiens*, and show similar patterns in DNA surprisal across exon-exon junctions, with a constant difference dependent on the species. All species show a sharp decrease in surprisal at position −10 and sharp increase at position −1, 1nt upstream of the exon junction, consistent with Fig. 2. These 9-mers are created by the juxtaposition of the last nucleotide of the 5′ exon with the first 8 nucleotides of the 3′ exon. An interesting observation is that the most frequent dinucleotide across the junction for both exon-exon and exon-intron boundaries is GG, suggesting that the joining of the G in position −1 with the following 8-mer is driving this peak. For coding mRNA most species show a tight and consistent pattern similar to that of DNA surprisal, however *D. melanogaster* is a clear outlier, with a different pattern and a significantly higher information content. This is likely to be in part a consequence of the increased proportion of coding sequence in *D. melanogaster* (22.1% versus < 3% for the other species in Fig. 3).

The sharp decrease in surprisal is at position −9 for coding mRNA surprisal. When computed at other values of k (Fig. S3), these minima shift with k, indicating that the nucleotides driving this pattern are at the end of the k-mer, and suggesting that this pattern arises from the lesser conserved exonic 3 ′ AG terminal motif. There is little variation surrounding splice junctions for amino acid sequences which is expected since the information content encoded at these positions is expected to be a feature of splicing information, which is ahead of the translation process. This is consistent for other values of k (Fig. S3) and similarly to Fig. S2, the coefficient of variation is reduced for larger k

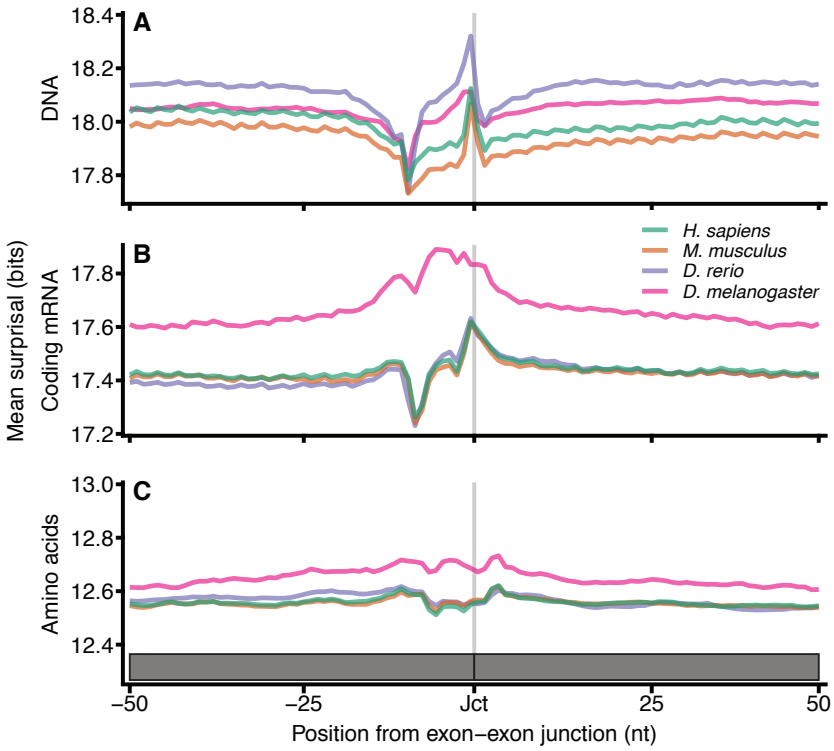

Figure 3 Surprisal across exon–exon junctions reveals non-coding information within coding sequences. For each of the 4 species, all 9-mers and amino acid 3-mers were extracted at each position relative to the exon-exon boundaries for all protein coding exons. The mean 9-mer DNA (A), 9-mer coding mRNA (B) and 3-mer amino acid (C) surprisal is plotted across the spliced exon boundaries centred on the exon junction. Data are plotted with the exon junction between positions −1 and 0 indicated by the grey line at position −0.5, and the *y* axes are a constant size, however the range is shifted dependent on the sequence type. In order to plot amino acid data on the same scale, the data for each exon junction were transformed into nucleotide space by repeating each amino surprisal value three times in succession. These data are in the positions by which they occur with respect to the exon junction in their respective phases. Source sequence data is as in Fig. 2, except that junctions within 100 nt of the translation start or end site were removed.

Full-size DOI: 10.7717/peerj.10063/fig-3

(Fig. S4). Together these data strongly suggest that codon redundancy allows additional signals to exist within the DNA without significant impact on the amino acid sequence encoded through the same space.

Effect of CG dinucleotides

The spectra of *k*-mers in the DNA sequences of several species were previously described by *Chor et al. (2009)*. Figure 4 shows the *H. sapiens* 12-mer DNA spectrum, which reveals three peaks in *k*-mer abundance corresponding to 12-mers occurring 4, 27, and 298 times in the genome. The spectrum can be viewed as three overlapping distributions, which are largely explained by the number of CG dinucleotides that occur within each 12-mer. This behaviour is not observed for any other dinucleotides (Fig. S5).

This CG dependency was also reported by *Chor et al. (2009)* for $k_{nt} = 8$, who suggested that it is a property of tetrapod genomes as a consequence of a global repression of CG

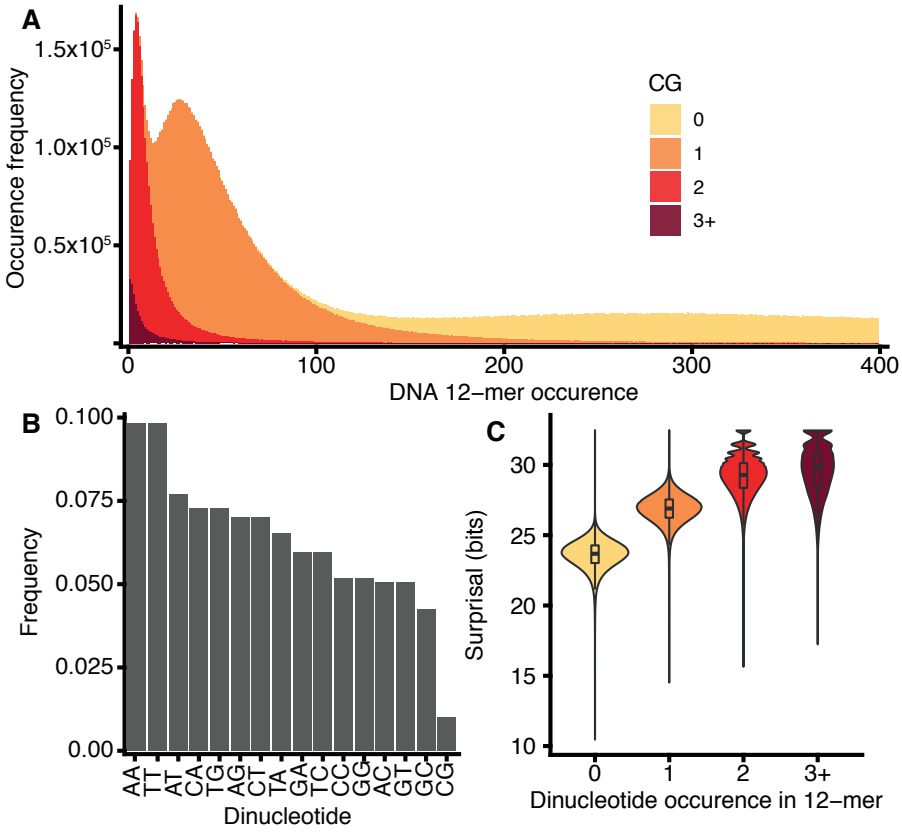

Figure 4 ***H. sapiens* DNA *k*-mer distributions are dependent on CG content.** (A) Stacked bar chart representing 12-mer occurrences for all 12-mers in *H. sapiens*, segregated according to the number of CG dinucleotides in each *k*-mer. *x*-axis: number of times a *k*-mer occurs in the genome, *y*-axis: number of distinct *k*-mers occurring at a given frequency. (B) The frequency of occurrence of the 16 dinucleotides in *H. sapiens* DNA normalised to the total number of all dinucleotides, including reverse complements. (C) Violin and box plot of surprisal distributions for 12-mers segregated by CG dinucleotide content.
Full-size DOI: 10.7717/peerj.10063/fig-4

dinucleotides, as shown in Fig. 4C for *H. sapiens*. This effect is further exacerbated by the concentration of CG dinucleotides at CpG islands, represented by a small number of CG-rich, common *k*-mers. The repression of CG dinucleotides is an established feature of genome evolution and results from the high mutation rate of CG dinucleotides caused by the deamination of methylated cytosines (reviewed by *Walsh & Xu (2006)*). A secondary consequence of the disproportionate rarity of CG dinucleotides is that *k*-mers that contain them tend to have high surprisal (Fig. 4D).

The frequency of CG dinucleotides in proximity to exon-intron junctions was investigated for *H. sapiens* DNA sequences (Fig. 5). Similar to Fig. 2, the 'valley' observed in mean surprisal centred at −16 is due to the effect of the polypyrimidine tract, shown explicitly in Fig. 5C. The CG dinucleotide frequency is 2-fold enriched within exons, providing a partial explanation for the increased information content in exons compared with introns. These results are in keeping with previous reports describing high CG differential between exon and intron regions (*Amit et al., 2012*). However, even within

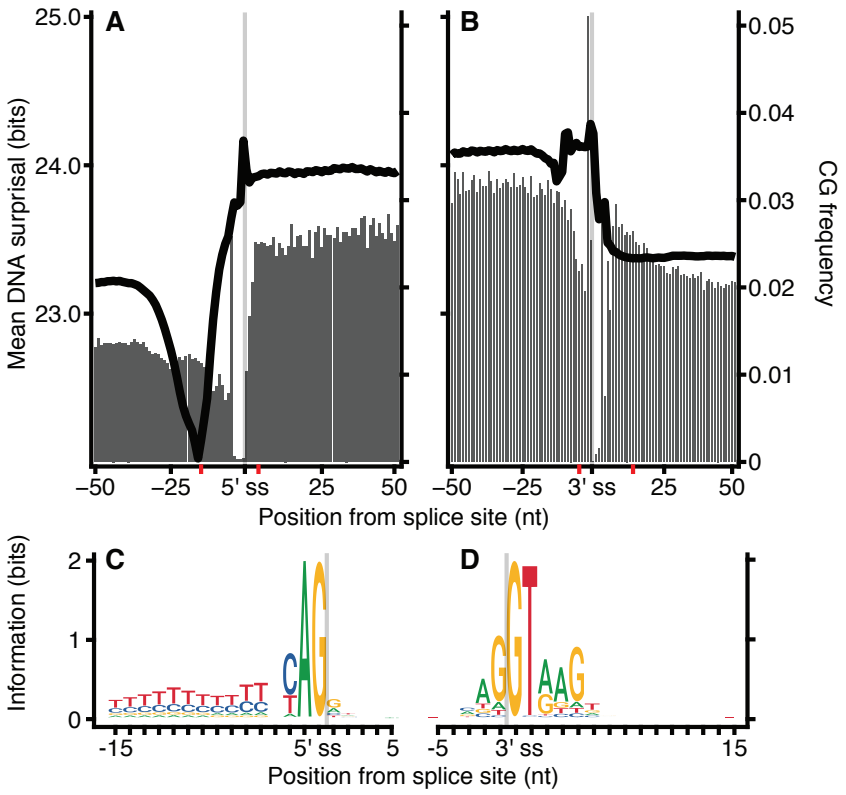

Figure 5 **CG dinucleotide count affects *H. sapiens* exon intron boundary surprisal.** (A, B) Average surprisal across exon-intron and intron-exon boundaries, as per Fig. 2 (line), and CG count frequency (bars). (C, D) sequence motifs at the 5′ and 3′ splice sites. The motif regions are identified in (A, B) by the red tick-marks on the *x*-axis.

Full-size DOI: 10.7717/peerj.10063/fig-5

exonic regions, the CG dinucleotide occurs at a much lower frequency than all other dinucleotides. No other dinucleotides show such a disproportionate difference between intron and exonic spaces (Fig. S6). However, CG dinucleotide usage is not sufficient to explain all patterns, as shown by the lack of correspondence between CG dinucleotide usage and surprisal across the polypyrimidine tract.

DISCUSSION

Here we describe a novel method for interrogating the genome using the self-information content of molecular sequences. The likelihood of occurrence of a *k*-mer is inversely correlated with surprisal, and hence information content. Genome-wide biological features are represented by sequences with lower surprisal because they occur more frequently and therefore tend to be encoded with common sequences. Conversely, loci with specific functions are encoded with rarer, higher surprisal sequences which contain more information. Our results also show that the CG dinucleotide is a major factor in *k*-mer surprisal. However, as shown in Fig. 5, while the CG dinucleotide is a major contributor to overall surprisal patterns, it is not the only factor at play.

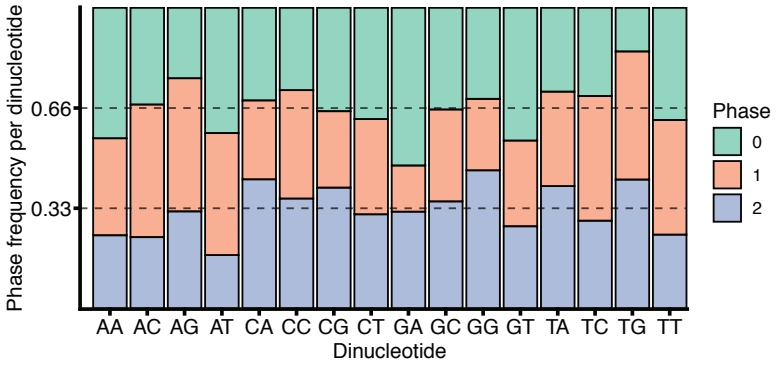

Figure 6 ***H. sapiens*** **phased dinucleotide usage.** Phase frequency of each dinucleotide within *H. sapiens* coding sequences split by the codon phase, phase 0 corresponds to the dinucleotide in codon positions 1 and 2, phase 1 would be codon positions 2 and 3 and phase 2 corresponds to position 3 in the upstream codon and position 1 in the downstream codon. Dotted lines are drawn at 1/3 and 2/3 representing unskewed dinucleotide phases.

Full-size DOI: 10.7717/peerj.10063/fig-6

The role of CG dinucleotides in CpG islands, DNA methylation (*Deaton & Bird, 2011*) and the hyper-mutability of methylated cytosines (*Misawa & Kikuno, 2009*) are all well understood, however the variation within the coding and non-coding regions around splice boundaries is less well characterised. While the increase in CG dinucleotide usage within coding regions, as observed in Fig. 5, may be due to the necessity to encode amino acids, a remarkable feature of the codon table is that all amino acids, and all amino acid sequences, can be represented without the use of the CG dinucleotide. Thus, all codons containing a C in position 2 have complete redundancy in the 3rd base. Further, while a CG in positions 1 and 2 (i.e., CGN) all encode arginine, arginine can also be encoded by AGY (where N corresponds to any nucleotide, and Y to a pyrimidine). Finally, all amino acids are redundant in the 3rd base between pyrimidines. It is therefore surprising that exon sequences are enriched for CG dinucleotides relative to the surrounding introns, particularly given the tendency for deamination driven C to T transitions. It is tempting to speculate that CG dinucleotide retention within exons is in part driven by the need to encode additional regulatory sequences within the same coding locus. This is in keeping with Fig. 6 which shows no substantial correspondence between CG usage and coding phase.

Our method has been applied genome-wide to regions involved in the established biological process of splicing. It successfully quantifies information patterns of known splicing motifs including splice sites and the polypyrimidine tract. The method can also be used to compare information content between different types of molecular sequence, such as coding mRNAs and amino acid sequences. Importantly, surprisal patterns observed at exon-exon boundaries for nucleotide sequences are not driven by the associated amino acid information (Fig. 3) suggesting that the position of exon boundaries can be accommodated by codon redundancy. This observation makes sense since translation occurs as a subsequent biological step after splicing.

The use of *k*-mers for sequence analysis is common and simple since implementations are available that are amenable to large scale computation. Here we use *k*-mers as the basis with which to compute surprisal patterns across loci. The method makes it possible to compare loci irrespective of the sequence type or similarity, as shown with the analysis of splice boundaries (Fig. 3). Having demonstrated the utility of our model, we will next apply it to the impact of SNVs in genetic diseases.

CONCLUSIONS

The information content of biological sequences has been modelled for many decades, but methods that rely on multiple sequence alignments are challenging when sequence divergence is large. Here, we describe a model in which a measure of the self-information content of molecular sub-sequences, *k*-mer surprisal, is used to quantify information content associated with biological features. Using splice sites as an exemplar, our model reveals clear patterns in surprisal around exon junctions that is observed consistently across a panel of evolutionarily diverse eukaryotes. Many of these patterns can be attributed to known biology, including binding motifs and patterns of dinucleotide usage. This surprisal model is complementary to existing approaches, and can be used to investigate sequence information content without the need for multiple sequence alignments.

CODE AVAILABILITY

The model was developed using MapReduce (http://mapreduce.sandia.gov) formulation in C++ (*Plimpton & Devine, 2011*). Analysis and figure plotting was performed in R using R-packages (*R Core Team, 2019*; *Dowle & Srinivasan, 2019*; *Wickham, 2017*; *Kassambara, 2020*; *Neuwirth, 2014*; *Wagih, 2017*; *Bengtsson, 2020*; *Lawrence, Gentleman & Carey, 2009*; *Wickham, 2011*; *Charif & Lobry, 2007*; *Wickham, 2019*). All code used in this work can be found at GitLab (https://gitlab.com/cruk-mi/genomic-kmer-surprisal-model).

ACKNOWLEDGEMENTS

We would like to thank Chang Sik Kim and Simon Pearce for their teaching contributions as well as insight during the development process.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work was funded by Cancer Research UK (CRUK) via core-funding to the CRUK Manchester Institute (grant no. A27412), the CRUK Manchester Centre (grant no. A25254) and the CRUK Beatson Institute (grant no. A17196). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
CRUK Manchester Institute: A27412.

CRUK Manchester Centre: A25254.
CRUK Beatson Institute: A17196.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Sam Humphrey conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Alastair Kerr, Magnus Rattray, Caroline Dive and Crispin J. Miller conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

All code is available at GitLab: https://gitlab.com/cruk-mi/genomic-kmer-surprisal-model.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10063#supplemental-information.

REFERENCES

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, Pupko T, Ast G. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports* **1**:P543–P556 DOI 10.1016/j.celrep.2012.03.013.

Bailey TL. 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27(12)**:1653–1659 DOI 10.1093/bioinformatics/btr261.

Bailey TL, Elkan C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology* **3**:21–29.

Bengtsson H. 2020. matrixStats: functions that apply to rows and columns of matrices (and to vectors). R package version 0.56.0. *Available at https://cran.r-project.org/web/packages/matrixStats/index.html*.

Berleant D, White M, Pierce E, Tudoreanu E, Boeszoermenyi A, Shtridelman Y, Macosko JC. 2009. The genetic code-More than just a table. *Cell Biochemistry and Biophysics* **55**:107–116 DOI 10.1007/s12013-009-9060-9.

Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics* **40**:1416–1425 DOI 10.1038/ng.264.

Charif D, Lobry J. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M, eds. *Structural approaches to sequence evolution: molecules, networks, populations. Biological and medical physics, biomedical engineering*. New York: Springer Verlag, 207–232.

Chor B, Horn D, Goldman N, Levy Y, Massingham T. 2009. Genomic DNA k-mer spectra: models and modalities. *Genome Biology* **10**:R108 DOI 10.1186/gb-2009-10-10-r108.

Cobb M. 2017. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology* DOI 10.1371/journal.pbio.2003243.

Corvelo A, Hallegger M, Smith CW, Eyras E. 2010. Genome-wide association between branch point properties and alternative splicing. *PLOS Computational Biology* **6(11)**:e1001016 DOI 10.1371/journal.pcbi.1001016.

Crick F. 1970. Central Dogma of Molecular Biology. *Nature* **227**:561–563 DOI 10.1038/227561a0.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugán JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P. 2019. Ensembl 2019. *Nucleic Acids Research* **47**:D475–D751 DOI 10.1093/nar/gky1113.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. *Methods in Enzymology* **91**:524–545 DOI 10.1016/S0076-6879(83)91049-2.

Dayhoff MO, Schwartz RM, Orcutt B. 1978. A model of evolutionary change in proteins. In: *Atlas of protein sequence and structure: supplement*. National Biomedical Research Foundation.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes and Development* **25**:1010–1022 DOI 10.1101/gad.2037511.

Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, Maulhardt T, Möhrmann L, Morstein J, Paffenholz SV, Röpenack P, Rückert T, Sandig L, Schell M, Steinmann A, Voss G, Wasmuth J, Weinberger ME, Wullenkord R. 2016. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, noncoding RNA and synonymous mutations. *EMBO Molecular Medicine* **8**:442–457 DOI 10.15252/emmm.201506055.

Dowle M, Srinivasan A. 2019. data.table: extension of 'data.frame'. R package version 1.12.8. *Available at https://cran.r-project.org/web/packages/data.table/index.html*.

Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. 2014. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Research* **42**:10681–10697 DOI 10.1093/nar/gku736.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Research* **32**:W187–W190 DOI 10.1093/nar/gkh393.

Gatlin LL. 1966. The information content of DNA. *Journal of Theoretical Biology* **10**:281–300 DOI 10.1016/0022-5193(66)90127-5.

Gibbs J. 1902. Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics. In: *Dover books on advanced science*. C. Scribner's Sons.

Hampikian G, Andersen T. 2007. Absent sequences: nullomers and primes. In: *Pacific symposium on biocomputing 2007, PSB 2007*.

Hartley RV. 1928. Transmission of information. *Bell System Technical Journal* **7**:535–563 DOI 10.1002/j.1538-7305.1928.tb01236.x.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89(22)**:10915–10919 DOI 10.1073/pnas.89.22.10915.

Herzel H, Ebeling W, Schmitt AO. 1994. Entropies of biosequences: the role of repeats. *Physical Review E* **50**:5061 DOI 10.1103/PhysRevE.50.5061.

Herzel H, Große I. 1995. Measuring correlations in symbol sequences. *Physica A: Statistical Mechanics and its Applications* **216**:518–542 DOI 10.1016/0378-4371(95)00104-F.

Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research* **17**:405–412 DOI 10.1101/gr.5987307.

Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. *Genome Research* **20**:1582–1589 DOI 10.1101/gr.105072.110.

Kassambara A. 2020. ggpubr: 'ggplot2' based publication ready plots. R package version 0.2.5. *Available at https://CRAN.R-project.org/package=ggpubr*.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Research* **21**:1360–1374 DOI 10.1101/gr.119628.110.

Koonin EV, Novozhilov AS. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**:99–111 DOI 10.1002/iub.146.

Koslicki D. 2011. Topological entropy of DNA sequences. *Bioinformatics* **27**:1061–1067 DOI 10.1093/bioinformatics/btr077.

Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**:1841–1842 DOI 10.1093/bioinformatics/btp328.

Lee Y, Rio DC. 2015. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annual Review of Biochemistry* **84**:291–323 DOI 10.1146/annurev-biochem-060614-034316.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences of the United States of America* **98(20)**:11193–11198 DOI 10.1073/pnas.201407298.

MacKay DJC. 2003. *Information theory, inference & learning algorithms*. New York: Cambridge University Press.

Maraia RJ, Iben JR. 2014. Different types of secondary information in the genetic code. *RNA* **20**:977–984 DOI 10.1261/rna.044115.113.

Matera AG, Wang Z. 2014. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology* **15**:108–121 DOI 10.1038/nrm3742.

Melnik SS, Usatenko OV. 2014. Entropy and long-range correlations in DNA sequences. *Computational Biology and Chemistry* **53(Part A)**:26–31 DOI 10.1016/j.compbiolchem.2014.08.006.

Misawa K, Kikuno RF. 2009. Evaluation of the effect of CpG hypermutability on human codon substitution. *Gene* **431**:18–22 DOI 10.1016/j.gene.2008.11.006.

Nemzer LR. 2017. Shannon information entropy in the canonical genetic code. *Journal of Theoretical Biology* **415**:158–170 DOI 10.1016/j.jtbi.2016.12.010.

Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. *Available at https://CRAN.R-project.org/package=RColorBrewer*.

Plimpton SJ, Devine KD. 2011. MapReduce in MPI for Large-scale graph algorithms. *Parallel Computing* **37**:610–632 DOI 10.1016/j.parco.2011.02.004.

Pritišanac I, Vernon RM, Moses AM, Forman Kay JD. 2019. Entropy and information within intrinsically disordered protein regions. *Entropy* **21**:662 DOI 10.3390/e21070662.

R Core Team. 2019. R: a language and environment for statistical computing. Vienna: R Foundation of Statistical Computing. *Available at https://www.R-project.org/*.

Raponi M, Kralovicova J, Copson E, Divina P, Eccles D, Johnson P, Baralle D, Vorechovsky I. 2011. Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. *Human Mutation* **32**:436–444 DOI 10.1002/humu.21458.

Saha K, England W, Fernandez MM, Biswas T, Spitale RC, Ghosh G. 2020. Structural disruption of exonic stem–loops immediately upstream of the intron regulates mammalian splicing. *Nucleic Acids Research* **48**:6294–6309 DOI 10.1093/nar/gkaa358.

Sahebi M, Hanafi MM, Van Wijnen AJ, Azizi P, Abiri R, Ashkani S, Taheri S. 2016. Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins. *Gene* **587**:107–119 DOI 10.1016/j.gene.2016.04.057.

Sanger F. 1952. The arrangement of amino acids in proteins. *Advances in Protein Chemistry* **7**:1–67 DOI 10.1016/S0065-3233(08)60017-0.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**:6097–6100 DOI 10.1093/nar/18.20.6097.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* **188**:415–431 DOI 10.1016/0022-2836(86)90165-8.

Schwartz S, Hall E, Ast G. 2009. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Research* **37**:W189–W192 DOI 10.1093/nar/gkp320.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**:379–423 DOI 10.1002/j.1538-7305.1948.tb01338.x.

Shi Y. 2017. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology* **18**:655–670 DOI 10.1038/nrm.2017.86.

Shreif Z, Striegel DA, Periwal V. 2015. The Jigsaw puzzle of sequence phenotype inference: piecing together Shannon entropy, importance sampling, and Empirical Bayes. *Journal of Theoretical Biology* **380**:399–413 DOI 10.1016/j.jtbi.2015.06.010.

Sickmier EA, Frato KE, Shen H, Paranawithana SR, Green MR, Kielkopf CL. 2006. Structural basis for polypyrimidine tract recognition by the essential Pre-mRNA splicing factor U2AF65. *Molecular Cell* **23**:49–59 DOI 10.1016/j.molcel.2006.05.025.

Tribus M. 1961. *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*. Van Nostrand, 649.

Vinga S. 2014. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics* **15**:376–389 DOI 10.1093/bib/bbt068.

Wagih O. 2017. ggseqlogo: a 'ggplot2' extension for drawing publication-ready sequence logos. R package version 0.1. *Available at https://cran.r-project.org/web/packages/ggseqlogo/index.html*.

Walsh CP, Xu GL. 2006. Cytosine methylation and DNA repair. In: *Current topics in microbiology and immunology*. 301. 283–315 DOI 10.1007/3-540-31390-7.11.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**:802–813 DOI 10.1261/rna.876308.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**:737–738 DOI 10.1038/171737a0.

Wickham H. 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software* **40(1)**:1–29.

Wickham H. 2017. tidyverse: easily install and load the 'Tidyverse'. R package version 1.2.1. *Available at https://cran.r-project.org/web/packages/tidyverse/index.html*.

Wickham H. 2019. stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. *Available at https://cran.r-project.org/web/packages/stringr/index.html*.

Wu J, Zhang Y, Mu Z. 2014. Predicting nucleosome positioning based on geometrically transformed tsallis entropy. *PLOS ONE* **9**:e109395 DOI 10.1371/journal.pone.0109395.

Yockey HP. 1974. An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology* **46**:369–406 DOI 10.1016/0022-5193(74)90005-8.

Yockey HP. 2000. Origin of life on earth and Shannon's theory of communication. *Computers and Chemistry* **24**:105–123 DOI 10.1016/S0097-8485(00)80010-8.

Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes and Development* **18**:1241–1250 DOI 10.1101/gad.1195304.