# Comparative Analysis of Splice Site Regions by Information Content

T. Shashi Rekha and Chanchal K. Mitra*

*Department of Biochemistry, University of Hyderabad, Hyderabad 500046, India.*

**We have applied concepts from information theory for a comparative analysis of donor (gt) and acceptor (ag) splice site regions in the genes of five different organisms by calculating their mutual information content (relative entropy) over a selected block of nucleotides. A similar pattern that the information content decreases as the block size increases was observed for both regions in all the organisms studied. This result suggests that the information required for splicing might be contained in the consensus of ∼6–8 nt at both regions. We assume from our study that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites, certain level of variability is still tolerated, which leads the splicing process to occur normally even if the extent of base pairing is not fully satisfied. We also suggest that this variability can be compensated by recognizing different splice sites with different spliceosomal factors.**

**Key words: splice site, substitution matrix, mutual information content, relative entropy**

## Introduction

Eukaryotes undergo the process of "RNA splicing", which involves the splicing of introns from heterogenous RNA (hnRNA or pre-mRNA) to form mature mRNA. Splice sites are characterized as donor (5′ boundary containing the dinucleotide GT in parent DNA or GU in pre-mRNA) or acceptor (3′ boundary containing the dinucleotide AG) regions. In addition to these dimers, a pyrimidine-rich region precedes AG at the acceptor site, and a short consensus follows GT at the donor site, while a very weak consensus appears at the branch point ∼30 nucleotides (nt) upstream of the acceptor site. A complex of nucleotide binding proteins and small nuclear RNAs (snRNAs), collectively known as the "spliceosome", recognizes these splice sites and excises introns by a concerted transesterification reaction (*1*). One important consequence of RNA splicing is that one gene can produce several different mRNA variants, or isoforms, simply by joining together different combinations of exons.

Several earlier studies have been reported for the detection of splice sites using different methods, such as the weight matrix model that uses the position compositional biases in splice sites (*2*). Artificial neural networks have been applied for the prediction of splice sites in different organisms with confidence levels better than previous methods (*3*). However, the reported results should be interpreted with caution as they were based on small datasets of limited number. A computational tool, GeneSplicer (*4*), was developed based on maximum dependence decomposition and performed better than previous tools. Recently the prediction of splice sites with dependency graphs and their expanded Bayesian networks has gained much importance because of its better performance (*5*). Current studies are being carried out to further understand and interpret the information contained in splice sites, as well as to develop a better method for their prediction with better specificity and sensitivity.

Detection of splice sites by using the two dinucleotides (GT/AG) is not meaningful because the frequency of these dinucleotides is very high in genes. Another important aspect to be considered is that the bases flanking them are also involved in the process and are expected to contain information required for splicing. Studying the consensus is also not directly useful, as they are highly variable not only within the species but also between species. Therefore, information theory comes to play a major role for the study of splice sites, which gives a quantitative measure of sequence conservation (or variability).

Information theory is an important tool (*6*) that has been often applied for understanding several key concepts in molecular biology (*7*). Information is defined as the amount of correlation between two random variables ($X$ and $Y$), which is measured as the

***Corresponding author.**
**E-mail: c_mitra@yahoo.com**

amount of entropy (uncertainty in a random variable) shared by them. This shared entropy is the information that one random variable contains about the other. It is a relative entity and is never absolute. In other words, mutual information is defined as a measure of the amount of information that one random variable contains about the other. It measures exactly the amount by which the entropy of $X$ or $Y$ is reduced by knowing the other, $Y$ or $X$ (*8*). This theory has gained much importance in biology by its applications to measure the information content of the nucleotide binding sites (*9*), identification of polymorphisms in DNA (*10*), prediction of RNA and protein secondary structures (*11*), prediction and analysis of molecular interactions (*12*), and drug design (*13*).

Study of horizontal correlations (between nucleotides along a sequence) is useful to identify features that can distinguish coding and non-coding regions in DNA (*14*). This gives the probability of finding nucleotides in the sequence that are correlated with each other. On the other hand, vertical correlations are important to find the probability of a nucleotide at a particular site by calculating the information content of the aligned set of sequences from its frequency of occurrence. Substitution matrices are thus useful to score these alignments perfectly.

Substitution matrix is a useful tool that scores the similarity between any two nucleic acid bases in terms of their ability to replace each other. By comparing a large number of similar sequences, one can obtain a matrix that describes the probability of a given nucleotide being substituted by another under the conditions of study. As probabilities are multiplicative, the logarithm is used to get an additive formulation. A number of techniques are now available for direct computations of substitution matrices, such as the BLOSUM (blocks substitution matrix) (*15*) and PAM (point accepted mutation) matrices (*16*). These matrices have been used extensively for global and local sequence alignments as well as database searches (*17*). They were also found to be significant for the study of core promoter regions (*18*).

Information theory has also been used for studying the features of spliceosome evolution and function (*19*). Studies have been carried out to correlate the intron length and the information content of the splice sites (*20*), suggesting that longer introns contain more information than shorter ones (*21*). Recently a comprehensive splice-site analysis using comparative genomics has been performed on different organisms by using the information content of the splice-site motifs, which proves that the identification of broad patterns in naturally-occurring splice sites, through the analysis of genomic datasets, provides mechanistic and evolutionary insights into pre-mRNA splicing (*22*).

It has become an important topic of research to characterize signals that govern the process of splicing in different organisms by information theory, which gives a broad idea about the distribution of information around the splice sites in different organisms. We have studied this aspect by carrying out a comparative analysis of donor and acceptor splice site regions in the genes of five different organisms (Table 1). We have constructed substitution matrices for the aligned set of sequences in the blocks of 6, 10, and 14 nt around the consensus dinucleotides (gt/ag) and calculated their information content, respectively (Figure 1). The substitution matrix specifically constructed for a given block is expected to work more efficiently than the one constructed for the whole genome sequences. In fact, we expect the difference to be evident among the three block databases. We have performed a broad analysis of the data distribution by calculating the information content at/around the splice sites, and achieved some interesting and informative results.

**Table 1 The Number of Genes and Splice Sites of the Five Organisms Studied***

| No. | Organism | No. of genes | Total No. of genes# | No. of splice sites | | Exon/intron boundaries |
|---|---|---|---|---|---|---|
| | | | | Donor | Acceptor | |
| 1 | *Arabidopsis thaliana* | 20,716 | 22,957 | 130,099 | 131,229 | gt-ag |
| 2 | *Caenorhabditis elegans* | 18,594 | 20,470 | 111,970 | 112,361 | gt-ag |
| 3 | *Drosophila melanogaster* | 10,612 | 15,624 | 72,737 | 73,167 | gt-ag |
| 4 | *Gallus gallus* | 16,567 | 16,568 | 168,120 | 169,990 | gt-ag |
| 5 | *Rattus norvegicus* | 19,146 | 19,197 | 181,782 | 183,476 | gt-ag |

*The splice sites with only "gt-ag" exon/intron boundaries were considered in our analysis. All other splice sites such as "gc-ag", "at-ac", and all the cryptic ones were excluded in the present study. However, we have included all the alternative splice sites in our analysis. #The total number of genes including alternative isoforms.

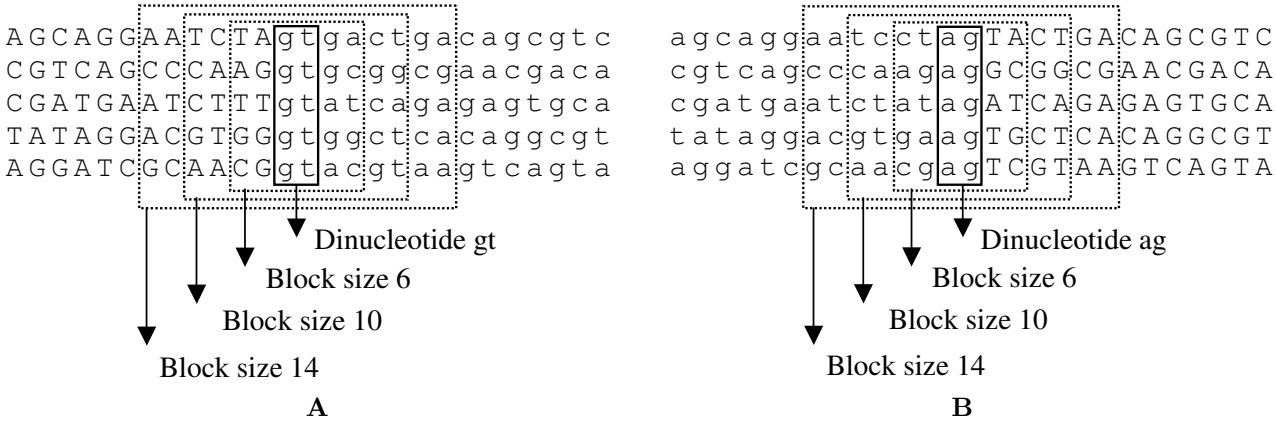


**Fig. 1** Illustrations of the construction of three different block databases for donor (**A**) and acceptor (**B**) splice sites. The splice sites are represented as donor (gt) and acceptor (ag) sites and the central dinucleotides (gt/ag) are aligned with 2, 4, or 6 nt taken on both sides. The three blocks are constructed for 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt, respectively. Note that the given sequences are for illustration only and are arbitrary. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters. The regions enclosed within the boxes are used for the computations of the substitution matrices.

## Results and Discussion

We calculated the mutual information content (relative entropy $H$) for each of the organisms studied from their log-odds matrices. The log-odds matrices scoring the alignments of the mononucleotide substitutions were obtained from the substitution matrices constructed for the frequency of occurrence of the nucleotide pairs. The information content values for the three blocks of all the five organisms studied are plotted as vertical box plots for both donor and acceptor sites (Figure 2). The 16 elements (4×4) of the $H$ matrix are plotted to get each box plot. These elements are the mean values of the given block and are directly comparable. Therefore, we are able to identify the contribution of the various elements individually.

The information content derived in this way is obviously a gross feature of the organism and perhaps can be divided into several groups such that the correlations within the groups are much more significant (compared to the whole genome; we expect the correlations between such groups may be quite less). The present plots in Figure 2 are more informative as they show a better distribution of the given data. We can clearly see the trends by following the median or the other percentiles. In all the plots we note that the 90 percentile bars are far from the median, suggesting that few points have relatively high values. The data points with high values were then examined manually and correlated with the particular elements of the $H_{ij}$ matrix as given in Table 2.

The box plots for the donor and acceptor sites of all the organisms studied (Figure 2) show interesting aspects that otherwise cannot be observed in the histograms (computed from the sum of $H_{ij}$ matrix elements) of the average mutual information content. We can see that the information content (the height of the box) decreases with the increasing block size for both donor and acceptor regions in all the organisms studied, suggesting that the distribution of nucleotides around the splice site junctions is more conserved (that is, the splice sites are more variable compared to the neighboring regions). The 6-nt block has the highest information content, and the information reduces considerably as we move away from the splice site. We speculate that the 6-nt block shows a greater variability (higher information content) and hence a higher selectivity. As we move to a larger window size, the variability decreases accordingly (as expected), suggesting that the selectivity of the spliceosomal binding is mainly dictated by the immediate neighborhood of the splice sites. This result reveals that the nucleotides of ∼2–3 nt flanking both sides of the splice sites are more important than longer distance nucleotides.

We also find that the median (50 percentile) values are more or less equal for all the plots. There exists a similar pattern of information content for both donor

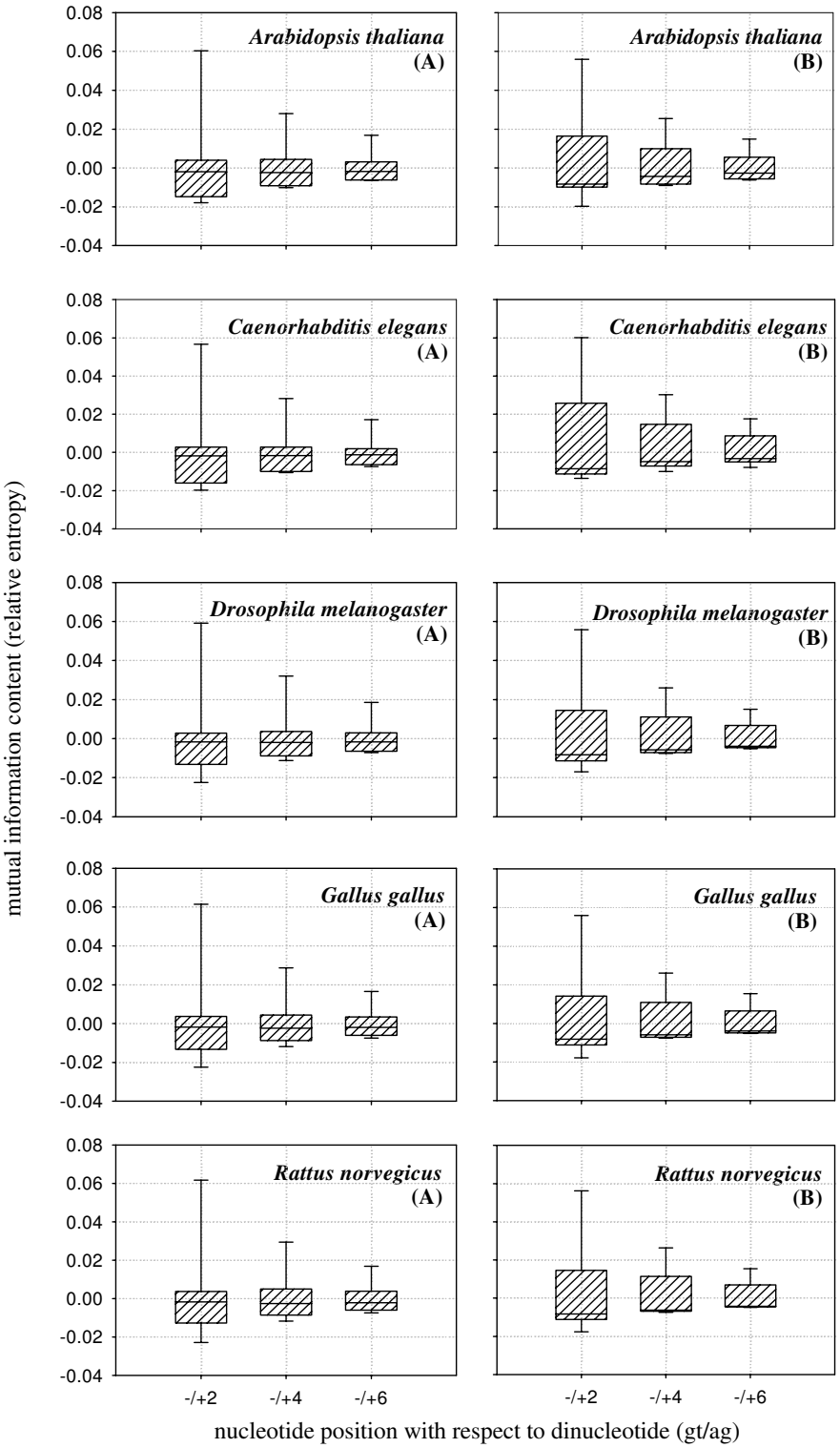


**Fig. 2** The mutual information content (relative entropy) calculated for donor (A; left column) and acceptor (B; right column) splice sites in the block sizes of 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt of the genes of five different organisms studied. The boundaries of the boxes represent the 25 (lower) and 75 (upper) percentile points. The horizontal line within the box represents the median value. The error bars show the 10 (bottom) and 90 (top) percentile points. It is clearly seen that the distribution is highly skewed and all the cases of the 90 percentile points are comparatively high in value. The median values show relatively little variation between the three blocks studied. All the graphs have been plotted on the same scale for ease in visual comparison.

**Table 2 Base Pair Preferences at Donor and Acceptor Splice Site Regions**

| Organism | Donor splice site region | | |
|---|---|---|---|
| | 6-nt block | 10-nt block | 14-nt block |
| *A. thaliana* | gg>tt>aa>ac>ca>cc | gg>tt>aa>cc>ac>ca | gg>tt>aa>cc |
| *C. elegans* | gg>tt>aa>cc>ca | gg>tt>aa>cc>ac>ca | gg>tt>aa>cc>ac>ca |
| *D. melanogaster* | tt>gg>aa>cc>ac>ca | gg>tt>aa>cc>ac>ca | gg>tt>aa>cc |
| *G. gallus* | tt>gg>aa>ac>ca>cc | gg>tt>aa>cc>ac>ca | gg>tt>aa>cc |
| *R. norvegicus* | tt>gg>aa>ac>ca>cc | gg>tt>aa>cc>ca>ac | gg>tt>aa>cc |
| | Acceptor splice site region | | |
| | 6-nt block | 10-nt block | 14-nt block |
| *A. thaliana* | gg>aa>cc>tt>ct>tc | gg>aa>tt>cc | gg>aa>tt>cc |
| *C. elegans* | gg>aa>cc>tt | tt>gg>aa>cc | tt>gg>aa>cc |
| *D. melanogaster* | gg>aa>cc>tt>ct>tc | gg>aa>tt>cc | gg>aa>tt>cc |
| *G. gallus* | gg>aa>tt>cc>ct>tc | gg>aa>tt>cc | gg>aa>tt>cc |
| *R. norvegicus* | gg>aa>cc>tt>ct>tc | gg>aa>tt>cc>ct>tc | gg>aa>tt>cc |

and acceptor sites in all the organisms studied, as they are equally significant for the binding of different spliceosomal proteins. We note that the values between 10–50 percentiles are very compact (less spread) while the values of 90 percentiles are far away from the median. This suggests that there are 1–2 values that are relatively high, which signify that the corresponding nucleotides are contributing to the high variability. In order to get a better understanding, we correlated the box plots of each organism with the individual elements of the $H$ matrix ($H_{ij}$, 4×4=16 individual values) to obtain the information about individual base pair preferences as given in Table 2.

## Donor (5′ splice site) region

We note from Table 2 that in the donor sequences the base pairs "gg" and "tt" have higher information content than "aa" and "cc" for all the cases. This is because the dinucleotide "gt" at the donor splice site is conserved and does not contribute to information content. Thus the high information content is attributed to the variability of the two nucleotides in the flanking regions of "gt", which suggests a high probability of each of the two nucleotides getting substituted by the other. The probability of adenine getting substituted by cytosine (or *vice versa*) is also significant. We can see from the 6-nt block of donor sites that guanine is more preferred in the flanking regions (1–2 nt) of "gt" in *A. thaliana* and *C. elegans*, while thymine is more preferred in the flanking regions of *D. melanogaster*, *G. gallus*, and *R. norvegicus*. We also see from Table 2 that the extent of variability decreases as the block size increases, suggesting that the nucleotides contributing to the variability are present in the neighborhood of the splice sites.

## Acceptor (3′ splice site) region

We also note that in the acceptor sequences the base pairs "gg" and "aa" have higher information content than "tt" and "cc" for most cases. This is due to the conservation of the dinucleotide "ag" at the acceptor site, which does not contribute to the information content. This observation suggests that the given nucleotides in the decreasing order of their preferences contribute to the variability in the consensus of these sites. In the flanking nucleotides of "ag", the probability of thymine getting substituted by cytosine (or *vice versa*) is also observed. We note that the consensus at the acceptor region is more conserved than that at the donor region as fewer substitutions are observed comparatively, which is also evident from the high information content observed for the 6-nt block (Figure 2). It also shows a decreasing order in the preference of nucleotides as the block size increases (Table 2). We note from the 10-nt and 14-nt blocks of acceptor sequences that thymine is more preferred in the flanking regions of "ag" in *C. elegans*, which is due to the presence of the short and highly conserved polypyrimidine tract that is adjacent to the acceptor splice site. The consensus sequence TTTTCAG/R at the 3′ end has been shown to be critical for its recognition and binding to the U2AF protein during the process of RNA splicing (*23*). All other organisms show general trends in the distribution of the nucleotides.

# Conclusion

We assume from these observations that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites (gt/ag), there still exists a certain level of variability in the consensus, signifying that some substitutions are found to be tolerable at certain positions. This is presumed to respond to the different spliceosomal factors that lead the splicing process to occur selectively. Our study suggests that the information required for RNA splicing is contained in the consensus of ∼6–8 nt at both donor and acceptor regions, which are important for the binding of spliceosomal proteins to the splice sites as expected.

We have developed our own block databases and applied the concepts of information theory for this analysis. Our study gives a broad idea about the distribution of nucleotides at/around the splice sites and also gives a comparative analysis of the consensus sequences at both donor and acceptor regions of the splice sites, which is significant for the process of splicing in terms of their sequence conservation or variability. We assume that our study can provide some insights towards understanding the information hidden at/around the splice sites that are important for the process of splicing to occur efficiently. We conclude that variability is essential for the selectivity of the splicing process whereas conservation is desirable to restrict the degree of variability.

# Materials and Methods

## Database

The Exon-Intron Database (EID) released in September 2005 (http://hsc.utoledo.edu/bioinfo/eid/index.html) was downloaded for the present study. This database was built in FASTA format by utilizing the data obtained from GenBank. It is a database of protein-coding intron-containing genes, which contains gene sequences of different organisms along with their alternative isoforms (*24*). The splice sites with only "gt-ag" exon/intron boundaries were considered in our analysis. All other splice sites such as "gc-ag", "at-ac", and all the cryptic ones were excluded in the present study. However, we have included all the alternative splice sites in our analysis. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters. We selected the gene sequences of five different organisms in order to have a broad data distribution, including *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves), and *Rattus norvegicus* (mammal). Table 1 gives the details of the number of gene sequences and splice sites analyzed in the present study. Our objective was to select a broad range of species but otherwise the selection may be considered arbitrary. Therefore the present study can be considered as "typical" or "representative" with a reasonably broad representation.

## Construction of block databases

The databases of splice sites containing the gene sequences of the given organisms were used for the construction of block databases. We developed three different databases for the donor (gt) and the acceptor (ag) splice sites respectively by aligning 2, 4, and 6 bases flanking on either side of the dinucleotides (-gt- and -ag-) for all the organisms being studied. Consequently, we constructed three blocks of 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt for each of the donor and the acceptor regions as illustrated in Figure 1. We have used the three different block sizes in order to have a comparative analysis of the conservation of bases at the splice sites, which are involved in the process of splicing. This is a better approach when compared to earlier studies, which gives a good understanding of the distribution of information around the splice sites. Scanning the nucleotides one by one with entropy would have been computationally expensive and the information obtained might have been disproportionately low. The blocks obtained were then used for the computations of the substitution matrix.

## Substitution matrix

We constructed substitution matrices for the aligned set of sequences of the given block sizes to calculate their mononucleotide substitutions (*15*). For the construction of each substitution matrix, we counted the number of matches and mismatches of each nucleotide type in each column between the first sequence and every other sequence present in the database. The same procedure was followed for every sequence in the database for all the columns present, and the values obtained were stored in a 4×4 frequency table, which gives the number of possible pairs of nucleotides in

the database. For a database with a width of $w$ nucleotides and a depth of $s$ sequences, $ws(s-1)/2$ nucleotide pairs can be obtained, giving the frequency of occurrence of each of the 10 (4+3+2+1) different nucleotide pairs in the database. Thus we obtained a 4×4 frequency table, with each of its elements being represented as $f_{ij}$. This table was further utilized for the calculation of log-odds matrix. In our case, $w$ is taken to be 6, 10, or 14, while $s$ depends on the particular organism (Table 1) studied.

## Log-odds matrix

Log-odds matrix is suitable to score alignments, in which the frequencies of the nucleotides in the aligned sequences are used to construct the substitution matrix. Log-odds values are calculated by taking a logarithm to base 2 ($\log_2$) of the ratio of the observed (target) probability to the expected (background) probability. The observed probability ($q_{ij}$) for each $ij$ pair is calculated as:

$$q_{ij} = f_{ij} \,/\, \sum_{i=1}^{4} \sum_{j=1}^{i} f_{ij}$$

Then, the probability of occurrence ($p_i$) of the $i^{\text{th}}$ nucleotide in an $ij$ pair is calculated as:

$$p_i = q_{ij} + \frac{1}{2} \sum_{j \neq i} q_{ij}$$

The expected probability ($e_{ij}$) for each $ij$ pair is then calculated as $e_{ij} = p_i p_j$ for $i = j$, and $e_{ij} = p_i p_j + p_j p_i = 2 p_i p_j$ for $i \neq j$. The likelihood or the odds ratio matrix for each $ij$ pair is calculated as the ratio of the observed probability to the expected probability: $q_{ij}/e_{ij}$, which gives the likelihood of occurrence of the nucleotides in pairs rather than by chance. The log-odds value of each $ij$ pair is calculated as the logarithm of the odds ratio ($s_{ij}$), which is given as: $s_{ij} = \log_2(q_{ij}/e_{ij})$.

## Mutual information content (relative entropy)

The entropy of a random variable is a measure of the uncertainty of the random variable. Thus, it measures the amount of information required on average to describe the random variable. The entropy $H(X)$ of a discrete random variable $X$ with the probability mass (or density) function $p(x)$ is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

where the logarithm is taken to the base 2 and the entropy is expressed in bits. The relative entropy is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy or the Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$$

Mutual information is defined as a measure of the amount of information that one random variable contains about the other. The mutual information $I(X;Y)$ of two random variables $X$ and $Y$ with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is given as the relative entropy between the joint distribution and the product distribution $p(x)p(y)$ (*25*):

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

We calculated the mutual information content for each block as the relative entropy $H$ of the observed (target) probability to the expected (background) probability:

$$H_{ij} = q_{ij} \times s_{ij} \quad \text{or} \quad H_{ij} = q_{ij} \times \log_2 \frac{q_{ij}}{e_{ij}}$$

which is the product of the observed probability ($q_{ij}$) and the log-odds ratio ($s_{ij}$). The relative entropy of a log-odds substitution matrix is its ability to distinguish true alignments from other alignments, which appear by chance. We did not take over the sum of all the elements of the $H$ matrix; instead, we plotted them as individual elements ($H_{ij}$) in the form of box plots.

## Presentation of results

Instead of using a conventional histogram to display the results, we chose a box plot that shows the 25 and 75 percentiles as the box boundaries (Figure 2). The median (rather than the mean) value is shown within the box as a solid line. The error bars are shown as the 10 and 90 percentiles. This representation of data is more informative and gives a simple view of the

distribution of the given data. All the plots were generated using the commercial software Sigmaplot 9.01 (Systat Software Inc., Richmond, USA).

## Acknowledgements

This work was supported by the Council of Scientific and Industrial Research, Government of India (Senior Research Fellowship to TSR).

### Authors' contributions

TSR carried out the computations. Both authors participated in the discussion of the results and the interpretation. Both authors read and approved the final manuscript.

### Competing interests

The authors have declared that no competing interests exist.

## References

1. Lewin, B. 2000. Nuclear splicing. In *Genes VII.* Oxford University Press, New York, USA.
2. Staden, R. 1984. Computational methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12: 505-519.
3. Brunak, S., *et al.* 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220: 49-65.
4. Pertea, M., *et al.* 2000. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29: 1185-1190.
5. Chen, T., *et al.* 2005. Prediction of splice sites with dependency graphs and their expanded Bayesian networks. *Bioinformatics* 21: 471-482.
6. Shannon, C.E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379-423, 623-656.
7. Adami, C. 2004. Information theory in molecular biology. *Phys. Life Rev.* 1: 3-22.
8. Durbin, R., *et al.* 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.
9. Schneider, T.D., *et al.* 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415-431.
10. Rogan, P.K and Schneider, T.D. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* 6: 74-76.
11. Giraud, B.G., *et al.* 1998. Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E* 58: 6312-6322.
12. Adami, C. and Thomson, S.W. 2005. Predicting protein-protein interactions from sequence data. In *The Chemical Theatre of Biological Systems. Proceedings of the International Beilstein Workshop* (eds. Hicks, M.G. and Kettner, C.). Logos Verlag, Berlin, Germany.
13. Adami, C. 2002. Combinatorial drug design augmented by information theory. *NASA Tech Briefs* 26: 52.
14. Rekha, T.S. and Mitra C.K. 2006. $1/f$ correlations in viral genomes—a Fast-Fourier Transformation (FFT) study. *Indian J. Biochem. Biophys.* 43: 137-142.
15. Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.
16. Dayhoff, M.O., *et al.* 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (ed. Dayhoff, M.O.), Vol.5, pp.345-352. National Biomedical Research Foundation, Washington DC, USA.
17. Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219: 555-565.
18. Reddy, D.A., *et al.* 2006. Comparative analysis of core promoter region: information content from mono and dinucleotide substitution matrices. *Comput. Biol. Chem.* 30: 58-62.
19. Stephens, R.M. and Schneider, T.D. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* 228: 1124-1136.
20. Fields, C. 1990. Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res.* 18: 1509-1512.
21. Mount, S.M., *et al.* 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20: 4255-4262.
22. Sheth, N., *et al.* 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34: 3955-3967.
23. Hollins, C., *et al.* 2005. U2AF binding selects for the high conservation of the *C. elegans* $3'$ splice site. *RNA* 11: 248-253.
24. Saxonov, S., *et al.* 2000. EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* 28: 185-190.
25. Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, USA.