

3-2009

Ab Initio Exon Definition Using an Information Theory-based Approach

Peter K. Rogan

University of Western Ontario, progan@uwo.ca

Follow this and additional works at: <https://ir.lib.uwo.ca/biochempub>



Part of the [Biochemistry Commons](#), and the [Computational Biology Commons](#)

Citation of this paper:

Rogan, Peter K., "Ab Initio Exon Definition Using an Information Theory-based Approach" (2009). *Biochemistry Publications*. 10.
<https://ir.lib.uwo.ca/biochempub/10>

Ab initio exon definition using an information theory-based approach

Peter K. Rogan, Ph.D.

Abstract— Transcribed exons in genes are joined together at donor and acceptor splice sites precisely and efficiently to generate mRNAs capable of being translated into proteins. The sequence variability in individual splice sites can be modeled using Shannon information theory. In the laboratory, the degree of individual splice site use is inferred from the structures of mRNAs and their relative abundance. These structures can be predicted using a bipartite information theory framework that is guided by current knowledge of biological mechanisms for exon recognition. We present the results of this analysis for the complete dataset of all expressed human exons.

Index Terms—Biological System Modeling, Genetics, Information Theory, Monte Carlo Methods

I. INTRODUCTION

Transcribed coding sequences are processed in mRNA by coordinately recognizing acceptor and donor splice sites across an exon, according to the exon definition hypothesis^{1,2}. The selection of splicing signal sequences is complex, involving exon and intron sequences, complementarity with small nuclear (sn) RNAs, RNA secondary structure and competition between splicesomal binding sites^{3,4,5}. The major spliceosomes contain snRNAs to guide recognition to constitutive donor and acceptor sites, which define exons^{6,7,8} (U1, U2 and U4-U6). U1 ribonuclear protein (snRNP) interacts with the donor (or 5') splice site^{9,10}, and U2 (and U6) snRNP with the acceptor (or 3') and lariat pre-mRNA branchpoint sites^{10,11}. Although both U1 and U2 base pair to mRNA, the complexes formed with human splice donors and acceptors vary in their stability because the duplexes are often mismatched¹².

The intrinsic stability of these interactions can be analyzed using information theory, which comprehensively and quantitatively models the thermodynamics of functional sequence variation^{13,14}. Information theory-based methods are based solely on experimentally validated sites, in contrast with training models using both true binding sites and non-binding sites¹⁵. The average information, $R_{sequence}$, connotes the overall conservation of a set of sites bound by the same recognizing protein(s), whereas particular binding sites that are members of this set are ranked by their individual information contents

(R_i)^{17,20}. Information theory-based models achieve very high sensitivity and specificity for detection of human donor and acceptor sites²⁶ (>98%). Further, changes in the affinity of a protein or protein complex for its cognate binding site can be estimated from differences in individual information contents between sites¹⁷. From the second law of thermodynamics, a 1 bit change in information corresponds to at least a 2-fold change in binding strength. Strong binding sites have R_i values $> R_{sequence}$, and weak sites are those with $R_i < R_{sequence}$. The zero coordinate on the R_i distribution can be understood from a thermodynamic viewpoint. In theory, R_i values > 0 correspond to true binding sites, since as entropy increases, energy is dissipated upon binding to the nucleic acid sequence. Selection of the most frequent base at each position of the information weight matrix $[R_i(b,l)]$ produces the consensus sequence, which is the upper bound of the distribution of R_i values.

In the present study, we model exon definition by minimizing entropy of a bipartite sequence pattern contained within human exon and flanking intron sequences. Our goal is to determine whether donor and acceptor splice sites may be concomitantly identified with this algorithm, analogous to the *in vivo* mechanism. Unlike other bioinformatic approaches for exon recognition¹⁸, this model does not require any of the hallmarks of protein translation in order to define the exon. A bipartite module consists of left- and right- motifs separated by an unspecified sequence that is recognized as a functional unit. We optimize the parameters that result in the maximum number of left and right motifs being defined as the acceptor and donor splice sites of a set of known exons.

II. METHODS

A. Algorithm

The bipartite *cis*-regulatory module is found by minimizing Shannon entropy over a set of unaligned sequences containing the two motifs separated by a gap of unspecified sequence (H)^{19,20}. This determines the total information content of the exon ($R_{i,total}$) for left and right motifs of different lengths for the gap lengths, d , which separate them. Each motif is represented by a different position weight matrix. The objective function (total information content, $R_{i,total}$) is:

$$R_{i,total} = R_i(left | d) + R_i(right | d) - g(d),$$

where

$$R_i(m | d) = \sum_{l=1}^{J_m} (E(H_{nb}) - H_m(l)), \quad m \in \{left, right\}$$

and

Manuscript received March 10, 2009. The author is appointed as a Canada Research Chair in Genome Bioinformatics at The University of Western Ontario. His research is supported by a Canada Foundation for Innovation Infrastructure grant. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca). He can be contacted at: MSB389, Schulich School of Medicine and Dentistry, London ON Canada N6A 5C1 (phone: (519)661-4255; email: progan@uwo.ca).

$E(H_{nb}) = \log_2 |D| - e(n)$, $D = \{A, C, G, T\}$. J_m is the width of motif m and $e(n)$ is a sample correction, $H_m(l)$ is the entropy for motif m at position l ¹⁴. The gap surprisal function, $g(d)$, is defined as $-\log(n(d)/n)$, and $n(d)$ is the number of sites with length d . $g(d)$ increases uncertainty, since it decreases the overall information content. Details of the Monte-carlo-based entropy minimization procedure used, comparisons with other methods, simulations and model performance have been presented previously¹⁹. The algorithms have been implemented and successfully applied to determine single block and bipartite motifs of a variety of prokaryotic and eukaryotic bifurcated binding sites²⁰. We have developed, *Partite*, a C++ program that implements the algorithm (available from the author). This software permits analysis of either one or both strands, either a single block or bipartite model, selection of either a uniform or non-uniform background distribution, either zero or one site per sequence, and an option to apply or exclude the gap surprisal penalty. Other parameters that can be specified include the minimum and maximum gap lengths, the method used to determine the minimum entropy alignment (eg. Monte Carlo estimation, Gibbs sampling, simulated annealing, or a genetic algorithm), the width of the left and right (or single block) motifs, the nucleotide length of sequences flanking the respective motifs, the number of pseudocounts, a temperature and/or cooling factor, and the number of Monte Carlo cycles performed. The program takes a single concatenated set of raw sequences as input, each designated with a separate header line.

B. Data

Models of internal exons containing both acceptor and donor sites were derived from human exon sequences extracted from NCBI Build 36.1²². Data were limited to validated, expressed genes in the manually-annotated, human Vega database²³. 153,506 interstitial exon sequences, each flanked by adjacent 100 nucleotide upstream and downstream intronic intervals, were downloaded with the Ensembl Biomart tool²⁴. Initial and terminal exons of genes lacking both donor and acceptor splice sites were excluded.

C. Analysis

We varied the lengths of sequences separating bipartite patterns, the respective motif lengths, the inclusion of the gap surprisal term, and the number of Monte Carlo cycles in each entropy minimization. The number of sequences used to derive each model was also varied, such that 20 randomly sampled datasets were assessed for each quantity of sequences analyzed. Phylogenetic trees of information weight matrices were produced using the UPGMA method for each of datasets comprising the same number of sequences, as a means of detecting potential sampling bias. The left and right motifs of the bipartite motifs for each model were ranked according to their Pearson's correlation E-values, ie. expected number of times that a similarity this strong would be observed by chance in a target database of random motifs²⁵ relative to a validated set of aligned, single site donor and acceptor $R_l(b,l)$ matrices¹².

III. RESULTS

Predicted bipartite information theory-based models for exon definition were compared with models of validated splice sites. The accuracy of these models was determined from the predicted locations of left and right motifs in known exons. Published models of acceptor and donor sites are 28 and 10 nucleotides in length, respectively^{12,26}. Previously, residual information above background levels was detected 3 nucleotides further upstream and 5 nucleotides further downstream of the coordinates defined by these matrices. The average length of internal exons is 97 nucleotides, and 99.6% of all exon lengths are below than 500 nucleotides. We therefore specified 500 nucleotides as the maximum gap size separating the bipartite motifs. The average information of models based on individual, validated splice acceptors ($n=108,079$) and donors ($n=111,772$) are 7.5 ± 3.4 bits and 6.7 ± 2.3 bits.

The bipartite model based on the most comprehensive dataset (153,506 exons), unexpectedly, did not produce high sensitivity or specificity for detection of acceptor and donor splice sites. While the motifs found are related to known splice sites and contain sequence elements found in those sites, the locations of these sequences often do not coincide with exon boundaries, although they usually overlap splice donor and acceptor sites. The polypyrimidine tract immediately upstream of the majority of U2-associated splice acceptor junctions is a conserved feature of nearly all of the models derived. The comprehensive exon set and several other models exhibit a preference for TG dinucleotide motif far upstream of the polypyrimidine tract at a level of conservation which significantly exceeds the bit content of these positions in true acceptors. This presence of this conserved sequence is not consistent with any previously described motif, including the branchpoint recognition site²⁷. The generally conserved AG dinucleotide that defines the exon-intron junction is absent from this model, indicating the the TG combined with the polypyrimidine submotifs together exhibit lower entropy than the natural splice site motif. The TG-submotif is separated from the polypyrimidine tract by 11 nucleotides. This suggested that the detection true positive acceptor sites might be enhanced by truncating the motif length, as other workers have done²⁸. This did not significantly improve either the accuracy of acceptor splice junction detection, but in some instances, the $R_l(b,l)$ matrices of these motifs exhibited greater similarity to natural acceptor sites.

Bipartite models based on fewer exons were found to be more accurate and were more highly correlated with the $R_l(b,l)$ matrices derived from known splice sites (Table 1). The best models were obtained by bootstrapping random sets of either 2000 or 4000 exons from the comprehensive dataset. Increasing the numbers of exons (6000,8000,10000,15000) produced models with left motifs that abrogated recognition of the conserved AG dinucleotide at the acceptor splice junction, even though this sequence is essentially invariant in all U2-associated splice sites (which comprise more than 99% of genomic exons; Fig 1). Models based on 1000 sites (or fewer) resembled natural splice sites, but generally did not detect

both donor and acceptor sites. The optimum minimal exon length was found to be 60 nucleotides, which corresponds closely to shortest natural exon lengths in these datasets. Regardless of the number exons aligned, models with longer minimal exon lengths (ie. 75-100 nucleotides) could accurately detect either acceptor or donor splice sites, but not both types of sites as an ensemble. Specifying longer intersite distances seemed to decrease Type I errors at the expense of an increase in Type II errors through the elimination of short exons. Inclusion of the gap surprisal term in the model, did consistently improve accuracy, however improvements were generally modest. Nevertheless, it was necessary to specify minimum threshold lengths in order to avoid detection of other sequence signals within transcripts unrelated to splicing (Fig. 2). As expected, very short inter-motif minimum distances produced accurate models of either acceptor or donor sites, but not both. The motifs detected adjacent to the constitutive splice site motifs were generally 6 to 8 nucleotides in length, had R_{sequence} values ranging from 4 to 8 bits, and tended to be uniformly distributed within an ~200 nucleotide interval circumscribing either natural donor or acceptor splice sites. Several of these short motifs appear to be similar to $R_i(b,l)$ matrices of binding sites recognized by highly expressed splicing regulatory proteins²⁹.

IV. DISCUSSION

The optimal bipartite model produced by entropy minimization accurately detected 90% of known exons and was derived from either 2000 or 4000 exon and intron sequences. While this level of accuracy is comparable or better than other available approaches^{30,31,32}, even higher sensitivities and specificities will be required to predict the structures of mRNA splice forms from primary genomic sequences. Constitutive splice site recognition can be modulated by the effects of adjacent splicing regulatory sequences³³. Incorporation of additional motifs derived from these sequences (exon and intron enhancer and silencer elements) may boost the accuracy of the bipartite models derived in this study. Because of their shorter length and lower levels of conservation, these regulatory sequences have lower overall information content and have higher multiplicity than constitutive splice sites. The combinatorial effects of individual regulatory sequences may be additive. Gap surprisal terms will be required to correct for the distances between these sequence elements and neighboring constitutive splice sites. More complex models containing these features will be required to accurately describe the multitude of abnormal splice forms produced by mutations that affect normal mRNA splicing.

The minimal entropy models based on larger numbers of sites often do not include the highly conserved nucleotides proximate to the acceptor splice junction. A common characteristic of the models based on >4000 sites is the increased conservation of the polypyrimidine tract. These tracts can vary considerably in length among different splice acceptor sites. The left motif probably represents a major subset of strong splice sites that is selected for by the model

which cumulatively contains more information than the highly conserved nucleotides close to the splice junction. Furthermore, variation in the distance between the conserved polypyrimidine elements and the conserved nucleotides proximate to the splice junction cannot be detected by the entropy minimization algorithm, such the the conservation at the acceptor splice junction is not preserved in the model. The failure to detect the conserved submotif adjacent to the junction cannot be mitigated by decreasing the motif length, which tends to find other conserved patterns among large sets of exon sequences.

Splice site recognition is a multistep process, coordinated by the action of both small RNAs and numerous proteins. This aspect of the biological mechanism raises the possibility the current ab initio approaches may be inadequate to catalog and quantify the strengths of multiple nucleotide motifs that are recognized in all exons. Assuming the goal is to develop models that can be applied for all interstitial exons, multipartite models which allow for variable length gaps both within individual splice sites as well as between them may be necessary to model exon definition.

V. CONCLUSION

Bipartite methods based on entropy minimization can frequently identify donor and acceptor splice sites at exon boundaries without prior alignment. Reasonably accurate models for exon definition can be obtained: (a) by limiting the number of exons aligned, (b) setting motif lengths to be comparable to those of known splicing signals, and (c) by specifying a minimum distances separating the motifs that is consistent with the distribution of known exon lengths.

ACKNOWLEDGMENT

The author would like to thank Dr. Thomas Schneider for valuable discussions.

REFERENCES

- [1] Berget, S. M. "Exon recognition in vertebrate splicing," *J. Biol. Chem.*, 270: 2411-2414, 1995.
- [2] Zheng ZM. "Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression," *J. Biomed Sci.* 11:278-94, 2004.
- [3] Moore, M. J., Query, C. C., Sharp, P. A in *The RNA World* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.), 303-358, 1993.
- [4] Craig, G. Simpson, G Clark, J.M. Lyon, J Watters, C McQuade, J Brown. *The Plant Journal*, 18(3), 292-302, 1999.
- [5] Will CL, Schneider C, MacMillan AM, Katopodis NF, Neubauer G, Wilm M, Luhrmann R, Query CC. "A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site," *EMBO J.* 20(16):4536-46, 2001.
- [6] Madhani, H. D., Guthrie, C, "Dynamic RNA-RNA interactions in the spliceosome," *Annu. Rev. Genet.* 28:1-26, 1994.
- [7] Steitz, J.A., Black, D.L., Gerke, V., Parker, K.A., Kramer, A. Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles. 115-154, 1988.
- [8] Seraphin, B., Kretzner, L., Rosbash, M. A. A U1 snRNA: pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site," *EMBO J.*, 7:2533-2538, 1988.
- [9] Zhuang, Y., Weiner, A.M. "A compensatory base change in U1 snRNA suppresses a 5' splice site mutation," *Cell*, 46 (1986), 827-835

- [10] Parker, R., Siliciano, P. G., Guthrie, C. "Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA," *Cell*, 49 (1987), 229-239.
- [11] Wu, J. A., Manley, J. L. "Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing," *Genes Dev.*, 3 (1989), 1553-1561
- [12] Rogan PK, Svojanovsky SR, Leeder JS. "Information theory-based analysis of CYP2C19, CYP2D6, and CYP3A5 splicing mutations," *Pharmacogenetics* 13:207-218, 2003.
- [13] Shannon, C.E. "Mathematical Theory of Communication," *Bell Systems Tech. J.* 27:379-623, 1948.
- [14] Schneider TD, G. D. Stormo, L. Gold, and A. Ehrenfeucht. "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, 188, 415-431, 1986.
- [15] Cawley S and Pachter L. "HMM sampling and applications to gene finding and alternative splicing," *Bioinformatics*, 19S:36-41, 2003.
- [16] Stephens R and T. D. Schneider. "Features of spliceome evolution and function inferred from an analysis of the information at human splice sites," *J. Mol. Biol.*, 228: 1124-1136, 1992
- [17] Schneider T. D. "Information content of individual genetic sequences," *J. Theor. Biol.* 189(4): 427-441, 1997.
- [18] Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, S. Brunak. "Splice site prediction in Arabidopsis thaliana DNA by combining local and global sequence information," *Nucleic Acids Res.* 24:3439-3452, 1996.
- [19] Bi C and PK Rogan, "Bipartite pattern discovery by entropy minimization-based multiple local alignment," *Nucleic Acids Res.* 32:4979-91, 2004.
- [20] Bi C and Rogan PK. "Information theory as a model of genomic sequences," in *Encyclopedia of Bioinformatics, Genomic, and Proteomics*, Wiley NY, 2005.
- [21] Bi C and Rogan PK, "BIPAD: a web server for modeling bipartite sequence elements," *BMC Bioinformatics* 7:76, 2006.
- [22] www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html#b36
- [23] Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. "The vertebrate genome annotation (Vega) database," *Nucleic Acids Res.* 36(Database issue):D753-60, 2008.
- [24] Hubbard TJP, BL Aken, S Ayling, et al. "Ensembl 2009," *Nucl. Acids Res.* 37: D690-D697, 2009.
- [25] Pietrovsky S. "Searching databases of conserved sequence regions by aligning protein multiple-alignments," *Nucleic Acids Res.* 24(21):4372, 1996.
- [26] Rogan PK, Faux B, Schneider TD, "Information analysis of human splice site mutations," *Hum. Mut.* 12:153-71, 1998.
- [27] Khan SG, Metin A, Gozukara E, Inui H, Shahlavi T, Muniz-Medina V, Baker CC, Ueda T, Aiken JR, Schneider TD, Kraemer KH. "Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk," *Hum Mol Genet.* 13:343-52, 2004.
- [28] Yeo G and Burge C, "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals," *J. Comput. Biol.* 11:377-94, 2004.
- [29] Nalla V and Rogan PK, "Automated splicing mutation analysis by information theory," *Hum. Mut.* 25:334-42, 2005.
- [30] Burge C and Karlin S. "Finding the genes in genomic DNA," *Curr Opin Struct Bio.* 8:346-354, 1998.
- [31] Lukashin AV and Borodovsky M. "GeneMark.hmm: new solutions for gene finding," *Nucleic Acids Res.* 26:1107-15, 1998.
- [32] Salzberg SL, Pertea M, Delcher A, Gardner M and Tettelin H. "Interpolated Markov models for eukaryotic gene finding," *Genomics*, 59:24-31, 1999.
- [33] Cartegni L, Chew S, Krainer AR, "Listening to silence and understanding nonsense: exonic mutations that affect splicing," *Nat Rev. Genet.* 3:285-98, 2002.

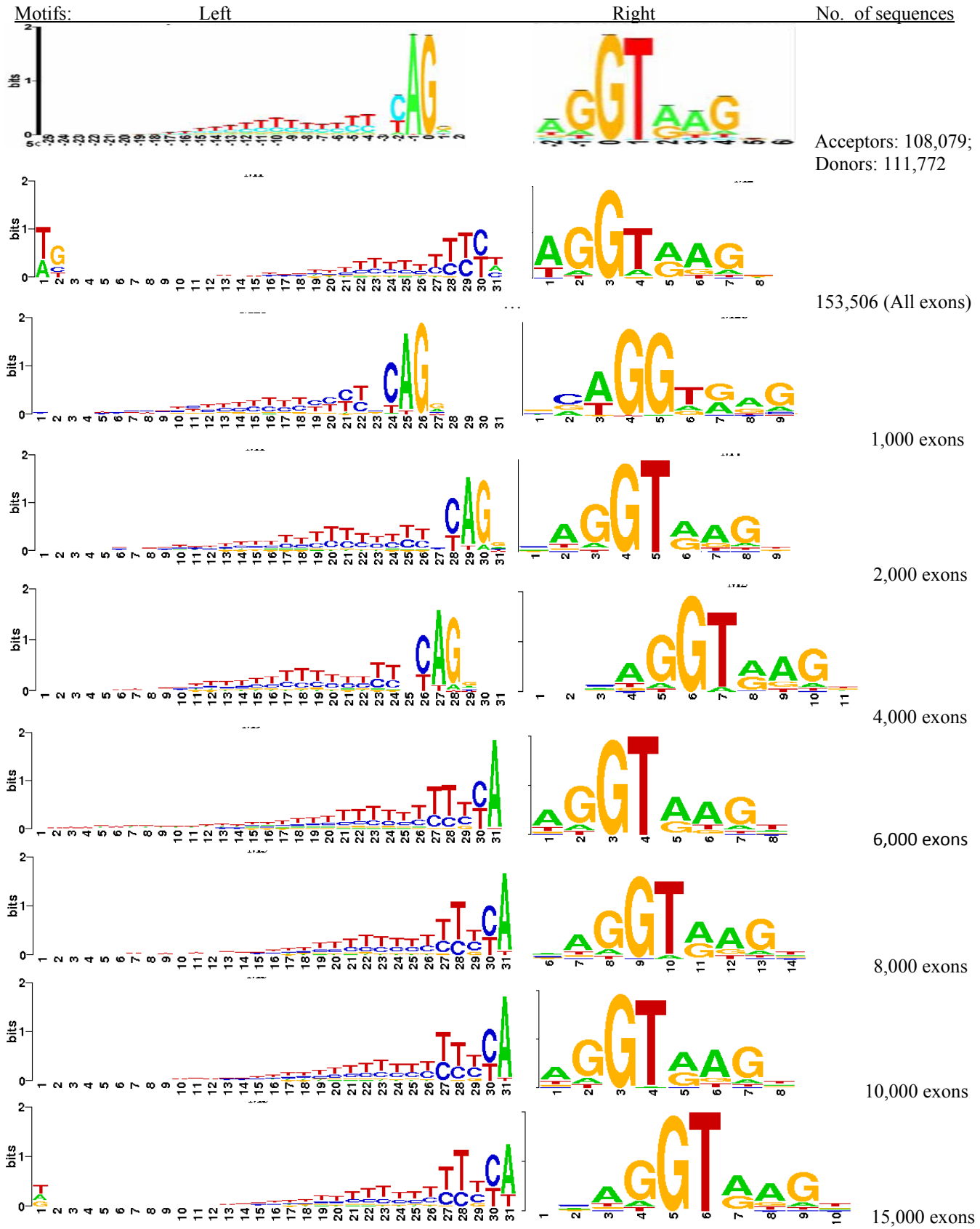


Figure 1. Representative sequence logos for exon definition models. The initial set of logos was derived by iterative refinement of acceptor (left motif) and donor (right motif) recognition sites from a genome-wide set (where all sites have $R_i > 0$ bits). All other sequence logos were derived as bipartite patterns with minimum and maximum intersite distances of 60 and 500 nucleotides,

respectively. The left and right motifs depicted here were specified to be 31 and 15 nucleotides in length, however other lengths and intersite distances were also evaluated (but are not shown).

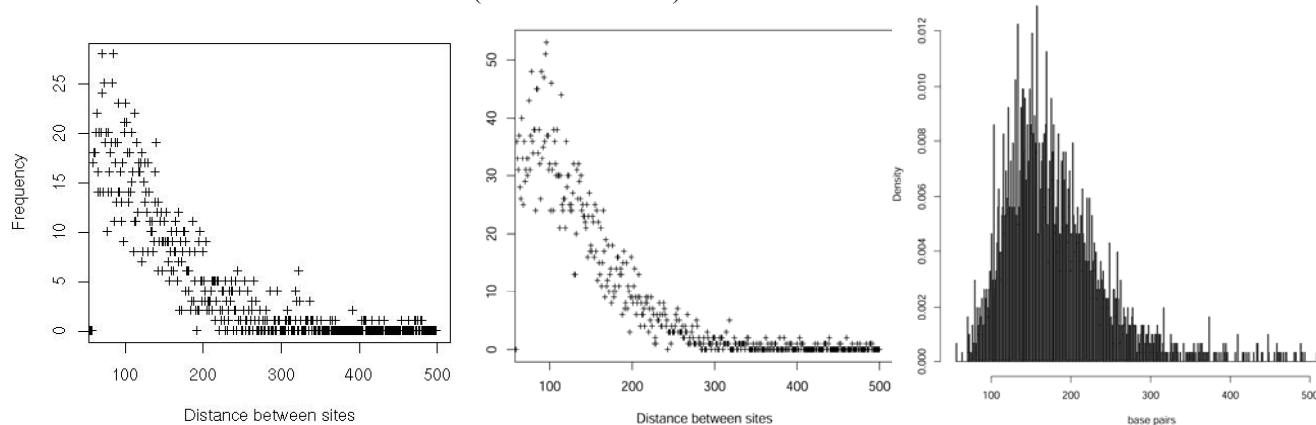


Figure 2. Predicted bipartite vs. natural exon lengths. Left and middle panels are the gap distribution of left and right motifs from 2000 and 4000 exons with flanking intron sequences. The right panel indicates the actual exon length distribution of the exons used to derive the predicted distribution in the middle panel.

Table 1. Properties of predicted bipartite models for exon definition

No.exons	Min Gap Length	Surprisal Applied	Left Length	Rseq (bits)	Accuracy	E-value ^a	Right Length	Rseq (bits)	Accuracy	E-value ^b	H (bits)
1000	60	-	31	10.7	0.88	5.8e-5	15	7.3	0.0	7.3e-1	73.9
2000	60	-	31	10.3	0.89	3.1e-7	15	7.9	0.86	3.5e-5	73.8
<i>2000</i>	<i>60</i>	+	<i>31</i>	<i>10.3</i>	<i>0.90</i>	<i>3.0e-7</i>	<i>15</i>	<i>8.0</i>	<i>0.92</i>	<i>3.8e-5</i>	<i>73.7</i>
2000	5	-	31	10.5	0.82	2.4e-5	10	7.6	0.0	6.6e-1	63.9
2000	5	-	10	8.1	0.0	7.4e-6	15	8.9	0.0	8.2e-1	33.1
4000	60	-	31	9.7	0.69	3.6e-7	15	8.1	0.70	3.2e-5	74.1
<i>4000</i>	<i>60</i>	+	<i>31</i>	<i>9.7</i>	<i>0.90</i>	<i>3.6e-7</i>	<i>15</i>	<i>8.1</i>	<i>0.99</i>	<i>3.2e-5</i>	<i>74.2</i>
4000	100	-	31	9.7	0.90	3.5e-7	15	6.5	0.31	6.9e-3	75.0
4000	1	-	31	10.0	0.88	2.1e-5	10	7.6	0.08	7.8e-1	64.4
4000	1	-	10	8.1	0.0	1.9e-1	15	8.4	0.99	3.2e-5	33.0
6000	60	-	31	10.0	0.0	2.9e-3	15	7.3	0.36	2.0e-2	75.1
6000	100	-	31	10.1	0.0	1.7e-1	15	6.9	0.99	2.9e-5	75.0
8000	60	-	31	9.3	0.71	1.3e-2	15	7.8	0.73	3.1e-5	74.9
8000	60	+	31	9.3	0.0	1.6e-3	15	7.8	0.93	4.0e-5	73.7
8000	100	-	31	9.3	0.0	1.8e-3	15	6.6	0.43	2.5e-1	76.1
8000	5	-	31	10.3	0.92	2.6e-7	10	7.3	0.0	7.1e-1	64.3
8000	1	-	31	10.3	0.87	2.7e-5	10	7.4	0.0	8.5e-1	64.3
8000	1	-	10	8.2	0.0	9.6e-6	15	8.4	0.0	7.3e-1	33.4
10000	60	-	31	9.4	0.0	2.2e-3	15	8.0	0.88	4.4e-5	74.8
10000	100	-	31	10.1	0.75	2.0e-7	15	6.4	0.21	4.4e-1	74.5
10000	100	+	31	9.5	0.66	1.3e-2	15	7.8	0.55	3.3e-5	74.5
10000	5	-	31	10.4	0.61	3.1e-7	10	7.4	0.0	6.4e-1	74.1
10000	1	-	10	7.9	0.66	1.7e-9	15	8.7	0.0	6.0e-1	33.1
15000	60	-	31	8.9	0.0	2.0e-2	15	8.1	0.95	3.0e-5	75.0
15000	100	-	31	9.3	0.0	1.7e-3	15	7.1	0.58	5.0e-5	75.9
153,506	60	-	31	8.7	0.0	1.1e-2	15	7.9	0.80	4.8e-5	75.5
153,506	100	-	31	8.7	0.0	1.4e-2	15	6.5	0.67	1.0e-4	76.9

Each row represents model parameters and averaged results for a randomized sample of datasets. Not all models are shown. The best fitting models are *italicized*. Comparisons are with validated acceptor^a and donor site^b models. No.exons: number of exons sampled from complete exon set; Min Gap Length: shortest distances between motifs. Surprisal applied: gap surprisal corrected; Left length: nucleotides in left motif; Right length: nucleotides in right motif; Rseq: R_{sequence} ; Accuracy: proportion of motifs which define true splice acceptor or donor site; E-value: chance probability, that there is another alignment with the splice site $R_i(b,l)$ with a similarity greater than the Pearson correlation coefficient; H: minimum entropy value for the alignment.