Mining conserved and divergent signals in 5' splicing site sequences.

Maximiliano Beckel[1,2†],Bruno Kaufman[1,2†], Marcelo Yanovsky[1,2], Ariel Chernomoretz[1,3*]

1 Fundación Instituto Leloir, Buenos Aires, Argentina
2 Instituto de Investigaciones Bioquímicas de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
3 Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Instituto de Física de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

†These authors contributed equally to this work.

Abstract

Despite the fact that the main steps of the splicing process are similar across eukaryotes, differences in splicing factors, gene architecture and sequence divergences in splicing signals suggest clade-specific features of splicing and its regulation.

In this work we study conserved and divergent signatures embedded in the sequence composition of eukaryotic 5´ splicing sites. We considered a regularized maximum entropy modeling framework to mine for non-trivial two-site correlations in donor sequences of 14 different eukaryote organisms. Our approach allowed us to accommodate and extend in a unified framework many of the regularities observed in previous works, like the relationship between the frequency of occurrence of natural sequences and the corresponding site´s strength, or the negative epistatic effects between exonic and intronic consensus sites. In addition, performing a systematic and comparative analysis of 5´ss we showed that lineage information could be traced not only from single-site frequencies but also from joint di-nucleotide probabilities of donor sequences. Noticeably, we could also identify specific two-site coupling patterns for plants and for animals and argue that these differences, in association with taxon-specific features involving U6 snRNP, could be the basis for differences in splicing regulation previously reported between these groups.

Introduction

In most eukaryotic genes coding DNA regions - called exons - are interrupted by non coding ones. These sequences, named introns, are co-transcriptionally spliced-out from the nascent transcript in a process called splicing [14, 17, 33]. RNA splicing results from a coordinated and sequential set of biochemical reactions that involve small nuclear ribonucleoproteins (snRNPs) which, together with less stably associated non-snRRP proteins, conform a dynamical molecular machinery called spliceosome [10, 33]. Two type of spliceosomes are known to operate in eukaryotes. A major U2 type (with U1, U2, U4, U5, and U6 snRNPs) that processes the majority of pre-mRNAs and a minor U12 type (with U11, U12, U4atac, U5, and U6atac snRNPs) that splices a minor fraction of

pre-mRNAs with so-called U12 type introns [35].

Despite some lineage-specific deviations, four major sequence cues serve to place the spliceosome at the right locations on the immature transcripts. For the vast majority of U2 spliceosomal introns, conserved GT and AG di-nucleotides are recognized at the beginning of an intron (5' splice site or donor site) and at the opposite intronic end (3' splice site or acceptor site) respectively. In addition, a branching point (BP), presenting a conserved A residue, is located 18 to 40 nucleotides upstream of the 3'ss. Finally, a poly-pyrimidine tract follows the BP and complete the necessary set of sequence cues used to guide the spliceosome assembly [11, 35, 38].

The 5' splicing site is a major requirement for splicing to take place. These donor sequences are involved in a key step of RNA splicing reactions in which boundaries between exons and introns are recognized. At this step, U1 identifies the 5' splice junction between adjacent exons and introns through a complex formation that depends on highly conserved base pairing between the 5' splice site and the 5' end of U1 snRNA. Specifically, U1 snRNA forms base pairs across intron-exon junctions, potentially base-pairing the last three positions of the exon (positions -3 to -1) and the first 6 positions of the intron (positions +1 to +6). In addition, before the first splicing catalytic step, U1 is replaced by U5 and U6 snRNPs, for which the snRNA also needs to bind to the exonic (-3 to +1) and intronic (+5,+6) portions of the 5'ss, respectively [1, 37, 38].

Given the relevance of donor sites, their sequence composition has been the subject of many studies which aimed to uncover non-trivial sequence patterns associated to relevant biology. In particular, information theoretical approaches were widely used to shed light on biological functionality and/or to trace the evolutionary history of splicing. For instance, in the earlies 90, Stephens & Schneider used information theory to analyze 1800 human splice sites. They could quantitatively estimate position-dependent contributions using a Shannon-like information measure and found that more than 80% of the sequence information (i.e. sequence conservation) was confined in the intronic part of donors sites [48]. The position-dependent information content measure was also considered by Sverdlov and collaborators to uncover differences in the way the information was distributed between exonic and intronic parts of nucleotide sequences in new (i.e. lineage-specific) and old (shared by two or more major eukaruote lineages) introns. They reported that old introns presented lower information content in the exonic than the intronic part of donor sites and, conversely, the opposite trend was true for new introns. This observation suggested an evolutionary splice signal migration from exon to intron during evolution [49]. More recently, Iwata & Gotoh compared 61 eukaryote species, and shown that even though donor site motifs resembled each other (suggesting that the spliceosome machinery is well conserved among eukaria) they did exhibit some degree of specificity [20].

Single-site statistics, like logos or consensus sequences, do not exhaust the relevant statistics embedded in 5'splicing sequences .There have also been attempts to characterize higher correlation patterns in donor site sequences beyond one-site statistics. One of the first results of this kind was reported in the aforementioned Stephens & Schneider´s paper. Despite that no mechanistic interpretation was provided, using information theory they were able to find significant mutual information values, of around 0.05, 0.07 and 0.04 bits, between human 5´ss positions (-2,+4), (-1,+5) and (-2,+5) respectively [48]. Almost ten years later, and following a completely different route, Thanaraj & Robinson considered decision trees to predict exon boundaries and found that long range dinucleotide associations (-1,+5) and (-2,+5) carried significant

splicing signals [50]. These observed couplings between (-1,+5) and (-2,+5) positions have also emerged in a human-mouse comparative genomic study carried on by Carmel and collaborators [5]. Two-point correlations in human 5´splice sites were also analyzed by Sahashi and collaborators. They reported that non-complementary nucleotides to U1 snRNA at specific positions were compensated by complementary nucleotides at other positions, suggesting that a stretch of complementary nucleotides either in an exonic or an intronic region was essential for proper splicing [41]. Denisov and collaborators presented evidence that supported and extended this idea. Through a comparative analysis of the genome of three mammals they found a well-defined pattern of epistatic interactions between the strength of nucleotides occupying different positions along the donor site. While the strength correlation within both, the intronic and the exonic part, were found to be positive (i.e. positive epistatsis), nucleotide strength correlations between intronic and exonic parts were found to present negative epistasis [13].

In this contribution we wanted to delve further into donor sequence regularities considering a maximum entropy approach in order to gain biological insights about splicing. Models rooted in the principle of maximum entropy have been applied to several biological problems in the last decade [3, 12]. A short list of applications includes, the study of the organization of natural flocks of birds [2] and neuronal activity levels [18, 39, 45, 51], reverse engineering of gene regulatory networks [26, 36], the study of transcription factors and DNA interactions [36, 42], and the analysis of protein structure and interactions guided by evolutionary information [15, 31, 47]. The working hypothesis behind the success of this modeling strategy is that, in all these systems, low-order (mainly pairwise) correlations play an important role, and a statistically accurate description can be established from observed probability distributions over pairs of elements [16, 39].

We relied on this approximation and used a maximum entropy approach to obtain a generative probabilistic model that allowed us to analyze sequence composition of donor sequences for 14 eukaryote organisms. With this modeling framework we were able not only to recover (and extend in scope) already observed patterns, but also to highlight novel regularities involving two-site interaction patterns.

The paper is organized as follow. We first present our modeling framework. We discuss how our regularized models accurately reproduce observed 1-site and 2-site nucleotide frequencies and serve to incrementally disentangle the hierarchy of coupling parameters sufficient to reproduce observed correlations at a given precision. Then, several aspects related with a data-driven energy function, naturally introduced in our maximum entropy approach, are discussed. We focus on the characterization of the identified coupling patterns afterwards. A comparative analysis allow us to identify robust and conserved signatures and extend previously detected epistatic signals in donor sites. Finally we present the analysis of divergent two-site interaction signals which highlights differences in the coupling patterns of plants and animals. We finally discuss some biologically relevant implications of our results and draw final conclusions in the last section of the manuscript.

Materials and Methods

Analyzed genomes

We considered annotated genomes from 14 eukaryotic species, including 3 plantae genomes (*Arabidopsis thaliana*, *Oryza sativa* and *Medicago truncatula*), 1 fungi genome (*Cryptococcus neoformans*) and, within the animal kingdom, 3 *Drosophila* genus genomes (*D. melanogaster*, *D. pseudoobscura*, and *D. yakuba*), 3 *Hominidae* family genomes

(*Gorilla gorilla*, *Pan troglodites* and *Homo sapiens*), 2 *Caenorhabditis* genus genomes (*C. elegans* and *C. briggsae*), and two more vertebrates, the zebrafish *Danio rerio* and the mouse *Mus musculus*. A custom script was used to automatically extract 9-base length donor sequences based on provided genome sequences (primary assembly FASTA files) and annotation (GTF files) downloaded from *Ensembl*. Supplementary table ST-S4 Table summarizes basic statistics for the considered 5'ss sequences (last three exon bases plus first six intronic bases) for each analyzed genome.

Statistical model

For each analyzed organism, we aimed to approximate the joint probability distribution function, $P(\vec{S})$, associated to the observed ensemble of $5'$ splicing sites. Each element of this ensemble was a 9-base long sequence (last three exon's positions followed by the first six intronic ones) $\vec{S} = (s_{-3}, s_{-2}, s_{-1}, s_1, s_2, s_3, s_4, s_5, s_6)$ where $s_i \in \{A, C, G, T\}$. We worked under the hypothesis that the sought distribution should be compatible with observed 1-site and 2-site marginal probabilities, $f_i(s_i), f_{ij}(s_i, s_j)$, and implemented a maximization entropy approach to find the minimal structured distribution consistent with these constraints. Under this framework the estimated density probability function,$\hat{P}(\vec{S})$, could be cast into a Boltzmann-like form (see [2,12] for technical details):

$$\hat{P}(\vec{S}) = \frac{1}{Z} e^{-E_d(\vec{S})} \tag{1}$$

with

$$E_d(\vec{S}) = -\sum_{i=1}^{9} h_i(s_i) - \sum_{i<j}^{9} J_{ij}(s_i, s_j) \tag{2}$$

playing the role of a data-driven energy. Z, also known as the *partition function*, can be considered here a normalization constant. $h_i(s_i)$'s and $J_{ij}(s_i, s_j)$ were the fitting parameters of our model. They conform single-site fields and pairwise-couplings that, together, make up the resulting energy score of a given sequence.

Overall, there were 36 single-site (4 bases for each of the 9 sites) and 576 two-site interaction (16 base combinations for 36 site-pairs) parameters that should be estimated in order to fulfill:

$$\sum_{\forall j \neq i} \sum_{s_j} \hat{P}(\vec{S}) = f_i(s_i) \tag{3}$$

$$\sum_{\forall k,l \neq i,j} \sum_{s_k, s_l} \hat{P}(\vec{S}) = f_{ij}(s_i, s_j) \tag{4}$$

Fitting procedure. Following [15] we implemented a regularized gradient descent scheme to fit the 612 parameters of our model. In each iteration we generated, using a Metropolis Monte Carlo procedure, an ensemble of 100000 sequences compatible with Eq 1 and the current model parameters. The fitting parameters were then updated according to the following rules:

$$h_i^{t+1} \leftarrow h_i^t - \epsilon_h [f_i(s_i) - f_i^m(s_i)] \tag{5}$$

if $J_{ij}^t = 0$

$$J_{ij}^{t+1} \leftarrow \begin{cases} 0 & \text{if } \mid \Delta \mid < \gamma \\ \epsilon_j [\Delta - sign(\Delta)] & \text{if } \mid \Delta \mid \geq \gamma \end{cases} \tag{6}$$

if $J_{ij}^t \neq 0$

$$J_{ij}^{t+1} \leftarrow \begin{cases} J_{ij}^t + \epsilon_j[\Delta - \gamma sign(J_{ij}^t)] & \text{if } \eta \geq 0 \\ 0 & \text{if } \eta < 0 \end{cases} \tag{7}$$

with $\Delta = f_{ij}(s_i, s_j) - f_{ij}^{model}(s_i, s_j)$ and $\eta = (J_{ij}^t + \epsilon_j[\Delta - \gamma sign(J_{ij}^t)]) * J_{ij}^t$.

It can be appreciated from these updating rules that the regularization parameter γ prevents non strong-enough coupling constants departing from zero.

Direct information. Direct Information is a scalar quantity originally introduced by Weigt and collaborators to measure the contribution of a given coupling constant, $J_{ij}(s_i, s_j)$, to the overall mutual information between sites (i, j) [54]. For a coupling constant of interest, $J_{ij}(s_i, s_j)$, the following joint probability function is considered:

$$P_{ij}^{DI}(s_i, s_j) = \frac{1}{Z_{ij}} e^{h_i^{DI}(s_i) + h_j^{DI}(s_j) + J_{ij}(s_i, s_j)} \tag{8}$$

where Z_{ij} is a normalization constant.

The single site fields $h_i^{DI}(s_i)$ and $h_j^{DI}(s_j)$ are determined in order to fullfill:

$$\sum_{s_j} P_{ij}^{DI}(s_i, s_j) = f_i(s_i) \tag{9}$$

$$\sum_{s_i} P_{ij}^{DI}(s_i, s_j) = f_j(s_j) \tag{10}$$

By construction, this two-site model can accurately reproduce 1-site empirical marginal distributions. Direct Information measures how much the dinucleotide probability distribution P_{ij}^{DI} differs from an independent site-model by considering a Kullback-Leibler estimation:

$$DI_{ij} = \sum_{s_i s_j} P_{ij}^{DI}(s_i, s_j) \log \frac{P_{ij}^{DI}(s_i, s_j)}{f_i(s_i) f_j(s_j)} \tag{11}$$

Dimerization energies

The dimerization free energy between the sequences of the donor sites and the complementary portion of the snRNA U1 is estimated using the RNAcofold program of the ViennaRNA 2.0 package [27], using default parameters.

Evolutionary rate

The evolutionary rate, ER, estimations were obtained from the OrthoDB site [23] for different phylogenetic clades: eukarya, matazoa, viridiplantae and fungi. ER measures whether a given Orthologous Group exhibits appreciably higher or lower levels of sequence divergence, derived from quantification of the relative divergence among their member genes. As stated in the OrthoDB site: 'ER´s are computed for each orthologous group as the average of inter-species identities normalized to the average identity of all inter-species best reciprocal hits, computed from pairwise alignments of protein sequences'. Slower and faster genome average evolution correspond to ER values larger and lower than one resepectively.

Results

The model

For each analyzed organism, we estimated a family of fitting models. For sequentially decreasing values of γ we got models of increasing complexity in a controlled way. With this fitting strategy we could disentangle the hierarchy of minimal sets of coupling constants J_{ij} necessary to adjust the observed two-site frequencies f_{ij} at a given accuracy level. We illustrated this point, for the *homo sapiens* case, in Fig 1.

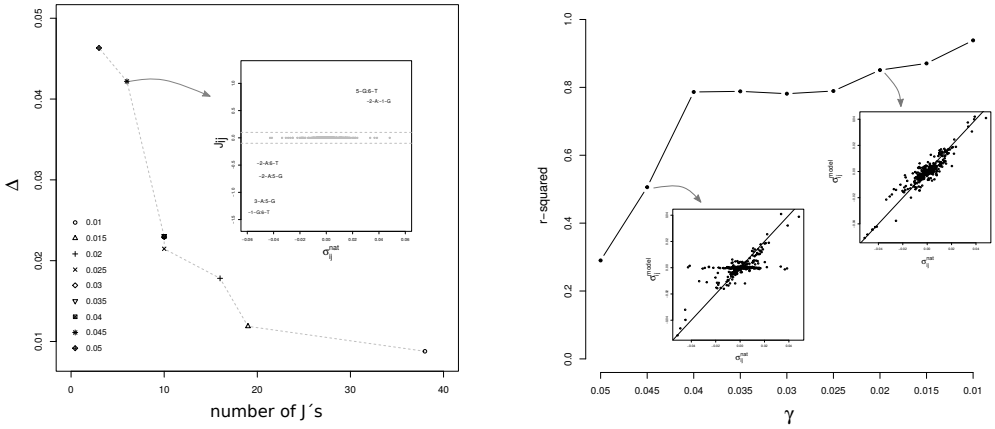


Figure 1. Role of regularization. Left panel: maximum absolute deviation between observed and estimated two site probabilities, $\Delta = max[abs(f_{ij} - P_{ij})]$, as a function of the number of the non-zero coupling constants identified at different regularization levels (different symbols, $\gamma \in [0.01, 0.05]$) for human donor sequence models. Inset: coupling constants $J_{ij}(s_i, s_j)$ as a function of observed two-site correlations $\sigma_{i,j}^{nat}$ for the $\gamma = 0.045$ model. Right panel: r^2 correlation coefficient between modeled and observed two-site correlations as a function of the regularization parameter γ. Insets: relationship between modeled, $\sigma_{i,j}^{model}$ and observed, $\sigma_{i,j}^{nat}$, two-site correlations at $\gamma = 0.045$ and $\gamma = 0.02$.

In the left panel of Fig 1 we show $\Delta = max[abs(f_{ij} - P_{ij})]$ - the maximum absolute deviation between observed, f_{ij}, and estimated, P_{ij}, two-site frequencies - as a function of the number of non-zero coupling parameters (the figure is parametrized by the corresponding regularization γ value). It can be observed that weaker regularization resulted in more complex models that fitted target frequencies increasingly well. We show in the inset the coupling parameter values, J_{ij}, as a function of the corresponding correlation observed in natural sequences, $\sigma_{ij} = f_{ij} - f_i f_j$, for the $\gamma = 0.045$ case. It can be appreciated that for this rather large regularization, the six non-trivial couplings recognized by this model were associated to the largest two-site correlations present in the data.

At every regularization level, the fitting models were able to generate an ensemble of sequences presenting one-site,P_i, and two-site frequencies P_{ij} that accurately reproduced the observed ones (R-squared > 0.99). In the right panel of Fig 1 we show the coefficient of determination (i.e. R-squared) of the regression of model-estimated against observed correlation values. It can be appreciated that with the incremental inclusion of more couplings (at lower regularization levels), the models could reproduce two-site correlations present in natural sequences increasingly well. We found a similar behavior for every analyzed genome (see Sup Fig S2 Fig). In particular, it can be seen from this figure that models obtained at $\gamma = 0.025$ regularization level presented around ten

coupling parameters different from zero and achieved a noticeable reduction in Δ. On the other hand, a leveling-off in Δ was consistently observed at $\gamma = 0.015$.

To better understand the roles and patterns of the fitting parameters, $\{h_i\}$ and $\{J_ij\}$, we show in Fig 2 a graphical summary of the $\gamma = 0.025$ model for human donor sequences.

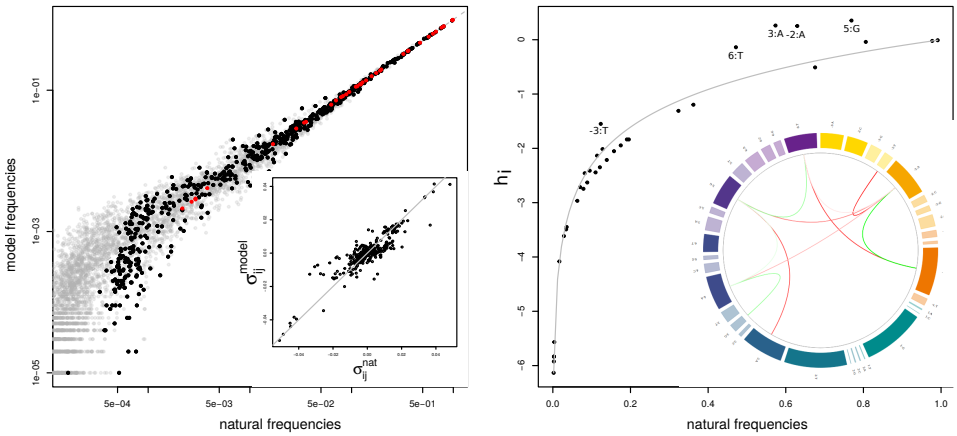


Figure 2. $\gamma = 0.025$ model frequencies and fitted parameters. Left panel: one-site, two-site and three-site model estimated frequencies (p_i, p_{ij}, p_{ijk}) as a function of the corresponding observed frequencies (f_i, f_{ij}, f_{ijk}), are depicted as red, black and gray dots respectively. Inset: relationship between model-estimated, σ_{ij}^{model}, and observed, σ_{ij}^{nat} two-site correlations. Right panel: $h_i(s_i)$ fitting values as a function of the observed one-site frequency, $f_i(s_i)$. The continuous gray line depicts the independent site model expected relationship $h_i^{indep}(s_i) \sim \log(f_i(s_i))$. Inset: circos representation of coupling interactions. The 36 circular ring boxes represent the 4 nucleotides per 9-site donor sequence. Warm colors were used for the three exonic sites whereas cold colors for intronic ones. The area of each box is proportional to the nucleotide-site obserevd probability $f_i(s_i)$. Positive and negative couplings are depicted with green and red lines respectively.

In the left panel of Fig 2 we show the relationship between estimated and observed one-site (red dots) and two-site (black dots) frequencies. The excellent agreement between natural and modeled quantities can be appreciated (R-squared > 0.99 for both cases). The inset of the panel shows the relationship between modeled and observed correlations. At a given level of regularization, correlations were a little bit harder to reproduce (R-squared~ 0.84). Noticeably, despite the fitting procedure just considered up to pair-wise observed marginals (i.e. f_i and f_{ij}) we found that the inclusion of a finite-set of pairwise couplings J_{ij} also allowed our model to reproduce higher-order statistics. In particular three-site probabilities (f_{ijk}, gray dots in Fig2) were very well captured in the simulated ensemble of sequences (R-squared $= 0.95$).

In the right panel of Fig 2 we show the h_i parameter values as a function of observed frequencies f_i. The estimated site-independent approximation for this quantities, $h_i \sim log(f_i)$, is depicted as a gray continuous reference line. The inset of the panel is a *circos* graphical representation of the suggested coupling connectivity pattern. The 36 circular ring boxes represent the 4 nucleotides per 9-site donor sequence. Hot colors were used for the three exonic sites, and cold colors for the six intronic ones. Four blocks of different

areas, representing the frequency of a given base $\{A, C, G, T\}$, were included for each site. Positive and negative couplings were depicted with green and red lines respectively.

From the figure it can be observed that many single-site parameters largely deviate from the independent-site approximation. This can be easily understood taking into account the presence of non-zero couplings, necessary to reproduce two-site statistics. Positive deviations occur whenever the site is involved in negative interactions and, vice-versa, negative deviations are expected when the site participates in stabilizing interactions. For instance, the single-site fields $h_5(G)$ and $h_6(T)$, that stabilized the consensus G and T bases at the last two intronic sites, presented larger values than the independent-site model. This happened because both sites participated in negative interactions with other consensus sites.

Data-driven energetics

According to our estimator of sequence probabilities introduced in Eq 1, the energy function defined in Eq 2 is a quantitative measure of how often a given sequence can be found along the entire genome. While low energy sequences correspond to the most prevalent splicing sites, high energy ones should be associated to rare and infrequent donor sequences.

Fig 3 shows the frequency distribution of donor sequence energies for the 502197 exon-intron annotated boundaries of the human genome. Ninety percent of sequences presented data-driven energy values laying in the $E_d \in [4.1, 9.7]$ energy range. On the low-energy side, the sequence of minimal energy, $\vec{S}^* = \{C, A, G, G, T, A, A, G, T\}$, exhibited perfect complementarity to the U1 RNA stretch, presenting an energy value of $E_d(\vec{S}^*) = 3.50$. This particular state was found to be the global minimum in the data-driven energy landscape of the entire ensemble of natural sequences (see Sup Mat S1 Appendix: Energy landscape). An ensemble of 2000 randomly generated sequencies was used to get a null reference for our data-driven energy scale. The gray dot and the horizontal black segment in Fig. 3 depict the mean and standard deviations of this null-model distribution ($E_d = 21.1 \pm 4.6$), a proxy for the completely disordered limit.

It can be appreciated that the energy distribution was slightly skewed toward high energy values. We noticed that high energy sequencies were enriched (gray tick marks in Fig. 3) in the 136 splicing sites that were reported to be targets of the minor U12 spliceosome according to the Intron Annotation and Orthology Database [32]. These findings suggested that correlations in our data-set were mainly driven by statistical regularities of the majoritarian U2 sequences. From our model´s point of view, U12 donor sites involved rather unusual sequences suggesting that they came from a different statistical distribution. Despite this observation we verified that the main findings (e.g. coupling patterns) and conclusions reported in this manuscript remained unaltered whether these sequences were removed from the training data-set or not. As we did not have information of the U12 sequences for the other analyzed organisms we decided to keep them in our data-set.

Fig. 3 also suggested that the data-driven energy (Eq.2) provided a 'natural' scale to characterize donor sites, going from completely ordered (perfect match against U1) to completely disordered sequences. Noticeably, the vast majority of exon-intron boundary sequences lied in between these extreme behaviors, in a region of natural variability where recognition is possible but a full binding with the recognition machinery is avoided. In addition, we found that our data-driven characterization nicely correlated with more physically sounded energy scales (inset of Fig 3). Not only the perfect complementary sequence to U1 (i.e. the one that minimized the binding energy) presented the minimal energy value but, more generally, an increasing monotone relationship could

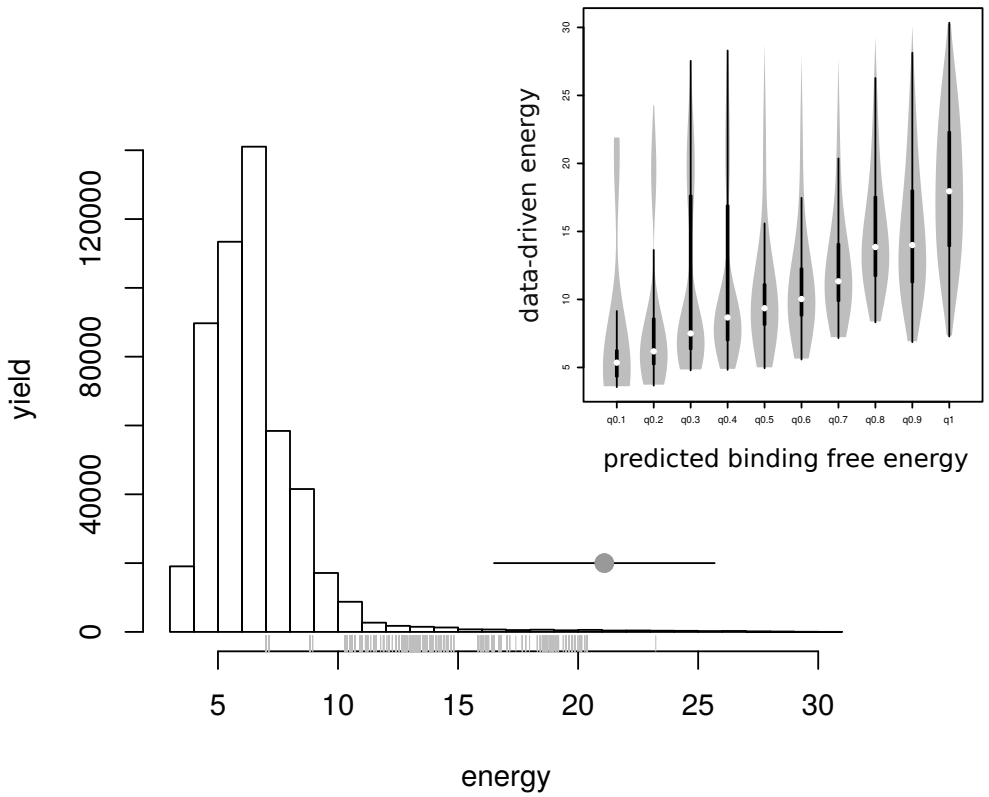


Figure 3. Energetics of human donor sequences ($\gamma = 0.025$). Distribution of data-driven energy values for the 502197 5'ss sequences of the human genome. Gray ticks were used to mark U12 sequence energies. The gray solid circle corresponds to the mean energy value observed for 2000 random 9-bases length sequences, whereas the black line shows the $\pm\sigma$ energy interval of this null distribution. Inset: boxplots of data-driven energy values estimated for 5'ss sequences of the ten deciles of the predicted dimerization energies against U1 smRNA.

be appreciated between our model sequence energies and estimations of the biochemical dimerization energy of 5'ss sequences against the U1 RNA stretch (see Methods).

Conserved coupling patterns

Our fitting procedure allowed us to identify sufficient sets of pairwise coupling interactions that could produce ensembles of sequences compatibles with two-site statistics presented in natural sequences.

Table 1 summarizes the average coupling strengths ($\gamma = 0.025$ models) between consensus (C) and non-consensus (NC) bases found at intronic (I) and exonic (E) sites for each analyzed organism. Despite some lineage specific peculiarities, this comparative analysis allowed us to identify a strong general trend in the pattern of pairwise interactions.

Stabilizing positive interactions could be recognized for consensus bases inside exonic (EC-EC) and intronic portions (IC-IC) of donors sites. Noticeably, a clear negative coupling between intronic and exonic consensus sites (IC-EC column in Table 1) could also be recognized across all the analyzed species. This last finding, suggesting that

Table 1. Conserved patterns ($\gamma = 0.025$). Mean interactions between different type of sites are shown for different organisms. EC, ENC, IC, and INC stand for exonic-consensus, exonic-non-consensus, intronic-consensus and intronic-non-consensus respectively.

Species	IC-EC	IC-ENC	INC-EC	INC-ENC	IC-IC	IC-INC	INC-INC	EC-EC	EC-ENC	ENC-ENC
cne	-0.08	0	0	0	0.34	0	0	0.92	0	0.17
ath	-0.37	0	0.02	0	0.19	0	0	0.91	0	0
mtr	-0.56	0	0	0	0.07	0	0	0.82	0	0
osa	-0.22	0	0	0	0.07	0.02	0.03	0.97	0	0
ptr	-0.22	0	0	0	0.59	0	0	0.78	0	0
ggo	-0.22	0	0	0	0.57	0	0	0.79	0	0
hsa	-0.55	0	0	0	0.16	0	0	0.82	0	0
dre	-0.34	0	0	0	0.24	0	0	0.91	0	0
mmu	-0.58	0	0	0	0.12	0	0	0.81	0	0
cel	-0.97	0	0.06	0	0.21	0	0	0	0	0.1
cbr	-0.92	0	0.05	0	0.38	0	0	0	0	0
dme	-0.97	0	0	0	0.26	0	0	0	0	0
dps	-0.66	0	0	0	0.75	0	0	0	0	0
dya	-0.8	0	0	0	0.6	0	0	0	0	0

simultaneous co-apparition of consensus sites both, at the intronic and exonic parts of donor sequences, was statistically disfavored, extended already presented evidence of negative epistatic signals found in mammals donor sequences [13, 41]. These results were also observed for more complex $\gamma = 0.015$ models (see Sup.Table S6 Table).

Divergent coupling signals

Despite the general high degree of conservation of the spliceosome, single site frequencies were shown to display a non-trivial degree of specificity in eukaryotes [20, 46]. We verified this in our data set but, in addition, we found that two-site statistics captured by our model, also exhibited this behavior. In panel A of Fig.4 we show a dendrogram (cos similarity, complete agglomeration method) considering two-site probabilities P_{ij} for $\gamma = 0.015$ models . Noticeably, the obtained hierarchical structure was identical to the one inferred from one site satatistics f_i.

It can be seen in this panel that the two-site statistical pattern of the fungus *cryptococcus neoformans* was the most divergent one, well separated from the one observed in plants and animals. Moreover a clear separation of the vertebrates, arthropods and nematodes organisms considered in this work, could be observed.

This finding suggested that non-trivial phylogenetic information was actually present in two-site correlations. In our model, they were captured at the level of coupling patterns which also reflected this tendency. For instance, it can be appreciated in Fig. 5 that circos interaction diagrams for $\gamma = 0.025$ models highlighted the existence of clear clade-specific patterns.

In order to further investigate the specificity of coupling signatures, we wondered if characteristic coupling patterns could be specifically associated to plants and animals. In order to explore this we consolidated, for one hand, the donor sequences of the three considered plants (*ath*, *mtr* and *osa* for a total of 423093 plant sequences) and, for the other, sequences coming from four representative animal species (*hsa*, *cel*, *dre* and *dme* for a total of 969705 sampled animal donor sequences). We then estimated maximum entropy models for each data-set (see Methods) in order to obtain representative coupling patterns for each analyzed group.

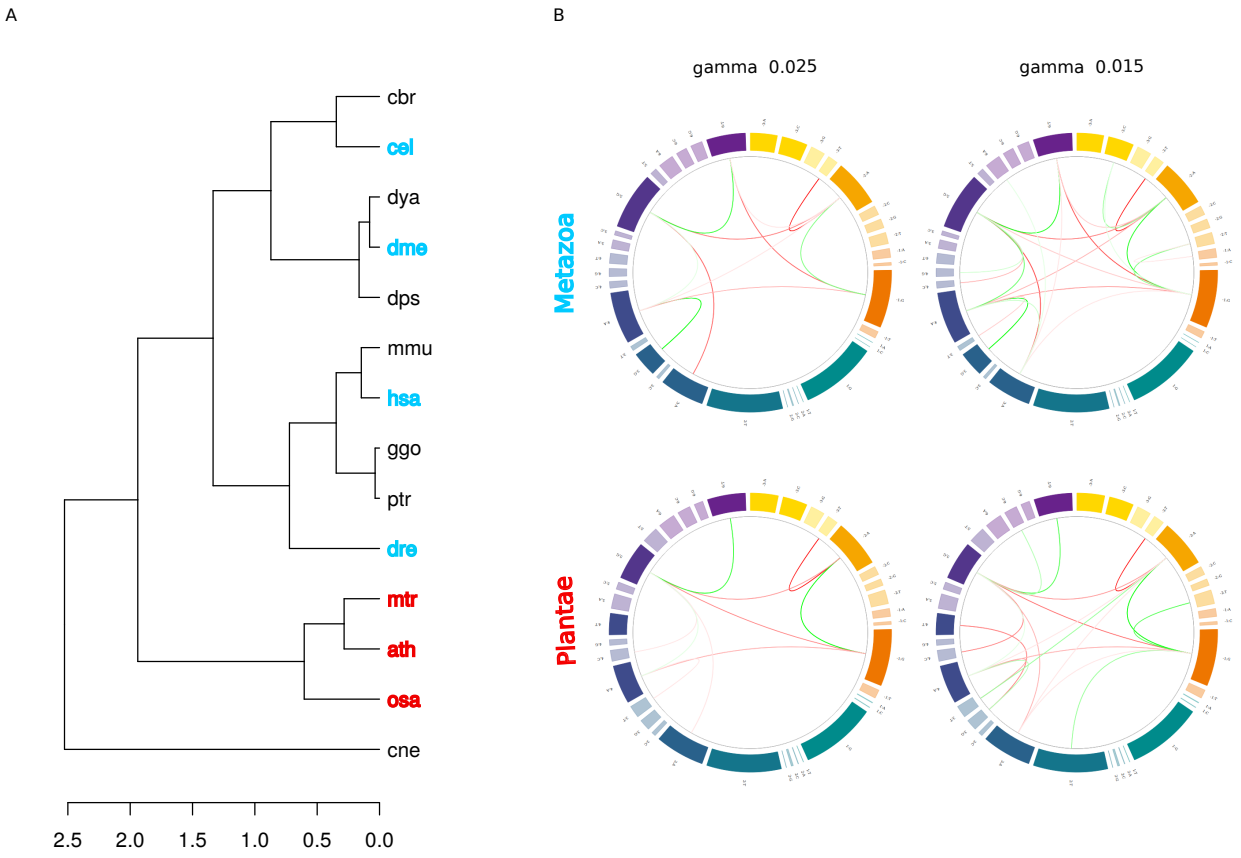


Figure 4. Divergent coupling patterns. A, Dendrogram grouping together species with similar two-site probabilities P_{ij}. Red and light-blue labels were used for organisms considered to analyze plant and animal donor sequences respectively. **B,** Circos diagrams for coupling parameters identified for metazoans (first row) and plantae (second row) donor sequences for $\gamma = 0.025$ (first column) and $\gamma = 0.015$ (second column) models.

The resulting coupling diagrams were shown in Fig.4. In the first column of panel B we included results from $\gamma = 0.025$ models. In this case, the rather strong regularization level served to highlight the main coupling pattern involved in the establishment of the most significant two-site correlations observed in each data-set (see Fig. 2 inset). Several interactions were shared by both groups. They involved positive couplings between consensus nucleotides (e.g. -2A:-1G, +5G:+6T) or negative interaction between consensus and non-consensus nucleotides (e.g. -3T:-2A) inside intronic or exonic parts of the sequence. At the same time, differences between both coupling patterns could also be appreciated. A strong positive interaction between consensus site +4A and +3G in the metazoan data-set was not detected in the plantae group. This stabilizing coupling could be understood in connection with the existence of a quasi-consensus site +3G in the metazoan group that was absent in the plant motif. In this way, despite being specific to one group, this kind of interaction was somehow expected from differences between groups taking place in 1-site frequencies.

Noticeably, we also found a divergent behavior for negative interactions involving the

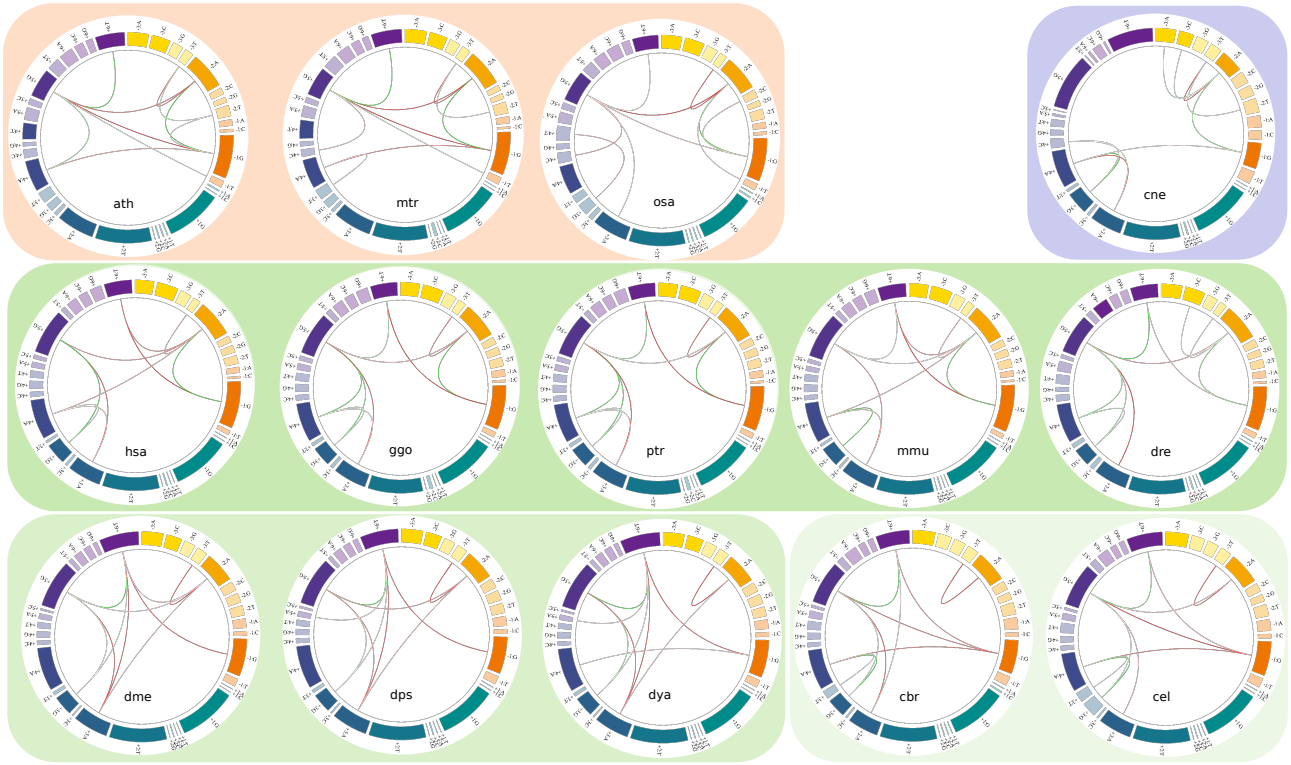


Figure 5. Pairwise interaction patterns ($\gamma = 0.025$). Pairwise interaction patterns obtained for each one of the 14 organisms studied in this work: *A. thaliana* (ath), *O. sativa* (osa), *M. truncatula* (mtr), *G. gorilla* (ggo), *P. troglodites* (ptr), *H. sapiens* (hsa), *D. rerio* (dre), *Mus musculus* (mmu), *D. melanogaster* (dme), *D. pseudoobscura* (dps), *D. yakuba* (dya), *C. elegans* (cel), *C. briggsae* (cbr), *C. neoformans* (cne). Plants are grouped together with a pink background (top left) whereas the sole fungus considered in this work, *cryptococcus neoformans*, with a purple background (top right). The rest of the patterns, with different shades of green, correspond to the animal kingdom. Vertebrates comprise the middle row, whereas the genus *drosophila* comprises the bottom left, and *caenorhabditis* the bottom right.

last two consensus intronic nucleotides +5G and +6T. A strong negative coupling was observed between +6T and -1G and between +5G and -3A nucleotides in the metazoan group, that was replaced by a negative interaction between +5G and -1G nucleoties in the plant group. These differences were not directly associated to 1-site statistics, and persisted in more complex models with relaxing regularization levels, like the one obtained at $\gamma = 0.015$ shown in the second column of Fig 4-B.

In order to quantitatively weigh the relevance of the effective couplings between donor sequence sites we relied on direct-information estimations (see Methods and Sup.Fig.S3 Fig). Statistically significant differences (Wilcoxon–Mann–Whitney test) were found between animal and plant groups for direct-information estimates involving site pairs: -2:+4 ($p = 0.02$), +3:+5 ($p = 0.007$), -2:+6 ($p = 0.007$), -1:+6 ($p = 0.007$) and +3:+6 ($p = 0.014$), (see Sup.Table S5 Table). These results supported the idea that major changes between plant and animals coupling patterns mainly involved the intronic last two nucleotide positions of donor sequences. We will go through a thorough discussion of possible implications of these findings in the next section.

Discussion

In this work we aimed to mine conserved and divergent signals in splicing donor sequences. We focused our attention on a set of 14 different eukaryotic organisms. The rationale of our approach was that 5´ donor sequences are a major cue for proper recognition of splicing sites, so statistical regularities of their sequence composition might reflect biological functionality and evolutionary history associated to splicing mechanisms.

Our entropy maximization strategy allowed us to recapitulate in a unified framework previous results and to gain new insights in regard to splicing. For instance, the data-driven energy scale E_d (Eq. 2), naturally accommodated the idea behind the SD-Score (defined as the logarithm of the frequency of a donor sequence) introduced by Sahashi and collaborators to predict the splicing outcomes observed in artificially designed minigenes [41]. In addition, E_d also correlated with estimated dimerization energies against U1 snRNA. Albeit that there are a large number of cis and trans elements that contribute to the regulation of the splicing process, this finding suggested that E_d by itself might reflect at some degree the strength of the donor site. In this sense, E_d not only served, by definition, to quantify how much a given sequence was represented along a genome, but at the same time acted as a meaningful scale to measure the degree of complementarity to U1.

Noticeably, we found that the majority of naturally occurring sequences presented intermediate E_d values (Fig. 3). The fact that a 9-nucleotide length perfect match was avoided agreed with previous observations stating that a minimal 5-6 Watson-Crick pairs were required for splicing site recognition, but too much pairing (> 7 bases) would be detrimental [5]. This loose binding might favor splicing reaction processivity and scenarios where the effective binding could be regulated by third players.

With the aid of our model we could also identify quite general two-site interaction patterns that suggested that this free-energy deficit was rooted in a non-trivial spatial distribution of matching pairs along splicing site sequences. On the one hand, despite some degree of taxonomic group specificity, couplings between consensus sites inside exonic and intronic parts of donor sequences were biased toward positive values (columns 5th and 8th of Table 1). On the other, for all the analyzed organisms, negative interactions were found between consensus nucleotides laying at different sides of the exon-intron boundary (first column of Table 1). These results extended previous observations obtained for human and mouse donor sequences [5, 13, 41] and supported the idea that a high complementarity level is alternatively favored either at the exonic or the intronic part of the splicing site.

Two-site couplings reflected that nucleotides of different positions were not independent. Some of the interactions detected for the different organisms analyzed in this contribution were already reported in the literature in the context of narrower studies focused only on human, mouse or some other mammalian genomes [5, 13, 41, 48, 50]. In fact, as far as we know, we showed for the first time that joint probabilities of nucleotide pairs carried biologically meaningful information in the sense that dendrograms inferred from them (which were identical to the ones obtained from one-site consensus motifs) closely followed phylogenetic relationships between the analyzed organisms (see Fig 4).

Our study allowed us to get a broader picture of two-site statistical regularities and, by doing so, it allowed us to identify subtle but statistically significant differences between coupling patterns in animals and plants. These differences mainly involved the last three intronic nucleotide sites of donor sequences (see Sup.Fig S3 Fig and Sup.Table S5 Table). In general, sites +4, +5 and +6, carried little information (0.22 bits and 0.35 bits mean information content in plants —ath, osa, mtr— and animals —hsa, cel, dme, dre— respectively) and showed relatively high level of variability. However, this does not necessarily means they lack biological importance. For instance, the relevance

of position +6 was already noticed by Carmel and collaborators in connection with splicing aberrations leading to familial dysautonomia [5]. In that work, the authors demonstrated that U1 snRNA base-pairing with positions +6 and −1 was a strong functional requirement for mRNA splicing of 5' splice sites.

Noticeably, positions +4 to +6 are known to base-pair not only with U1, but also with U6 snRNA during the splicing process. The conserved "ACAGA-box" sequence of U6 base pairs to the intron 5' splice site in the catalytically active spliceosome [22, 43, 44]. U6 presents a set of peculiarities that distinguish it from the rest of the spliceosome complexes. The U6 snRNA is the only one that is synthesized by pol III and that is not exported to the nucleus for the assembly of the complex. At the same time, the U6 snRNA is associated with a heteroheptameric complex of the LSm family of proteins, unlike the rest of the snRNAs that are associated with a complex of the Sm family, which arose from a duplication event of LSm proteins [52]. LSm are grouped into two complexes: LSm 1-7, mainly involved in the decapping process [4, 40] , and LSm 2-8, which makes up the core of U6 [24, 28, 30]. These proteins have been related to multiple regulatory functions, being considered in plants as a key piece in the adaptation and response to environmental changes [7, 19]. Having to interact with two different components of the spliceosome, these donor sites might present characteristic patterns of specificity that echo evolutionary divergent processes [46].

Several differences were already highlighted between plants and animals in regard to splicing mechanisms. Arguably, the most straightforward ones involve large differences reported in typical intron lengths and the prevalence of different kind of splicing events: exon skipping for animals and intron retention for plants [9, 35].

In addition, many recent contributions highlighted not only gene-architectural but also functional differences between alternative splicing taking place in animals and plants [8, 29, 34, 53]. In particular, the prevalence of intron retention events coupled to NMD transcript degradation and nuclear sequestration suggested that, differently from animal organisms, splicing in plants could play an explicit functional regulatory role rather than act expanding proteomic diversity [6, 9, 21, 25, 35]. In this context, our results highlighted significant differences in two-site interactions involving donor site nucleotide positions relevant at functional and evolutionary levels. Based on our findings we believe that the specific connection of these differences with distinctive mechanistic features of splicing processes in plant and animal organisms deserves to be further investigated.

Conclusion

In this work we followed a maximum entropy program to obtain regularized probabilistic generative models of donor sequences for 14 different eukaryote organisms.

Our approach allowed us to introduce a data-driven energy scale that reflected the abundance of a given sequence along a specific genome. Noticeably this energy statistic correlated with binding free-energy estimations against the U1 smRNA sequence and, at the same time, provided a sensible scale to characterize a given sequence in connection with the idea of completely ordered and/or completely disordered sequence states.

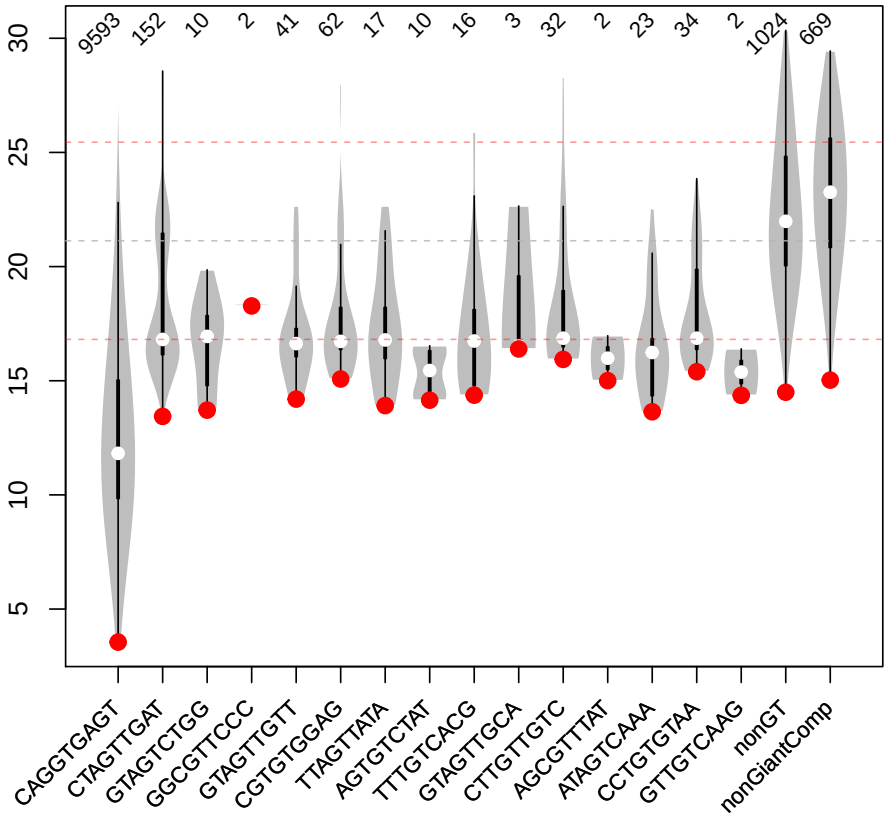
In our work we also showed that joint di-nucleotide probabilities of donor sequences carry lineage information. In fact, with the aid of our models, we could identify minimal sets of coupling patterns that could reproduce, at a given regularization level, observed two-site frequencies in donor sequences. The analysis of these interactions across organisms allowed us to identify specific two-site coupling patterns differentiating plants and animals. This sequence composition signature was embedded in two-site interactions involving the last nucleotides of the intronic part of donor sequences, which suggested that they could be related to taxon-specific features of the transcript interaction with

U6 snRNP.

Supporting Information

S1 Appendix: Energy landscape. To further characterize the energy landscape of donor sequences, we analized the $\gamma = 0.025$ model for human donor sequences considering a network graph approach. Representing individual sequences with nodes, we connected two of them if they were one mutation away (i.e. Hamming distance of 1). The edges were directed in a single direction, going from the higher-energy node to its lower-energy counterpart.

In this directed network, we found 243 local-minima sequences associated to nodes with no outwards edges (i.e. null out-degree). For each one of these attractor-nodes we could estimate the in-ward connected component, i.e. the set of nodes that could reach a given local minimum-energy node in a finite number of steps. These sets of nodes were a proxy of the extent of the basin of attraction of a given attractor sequence.



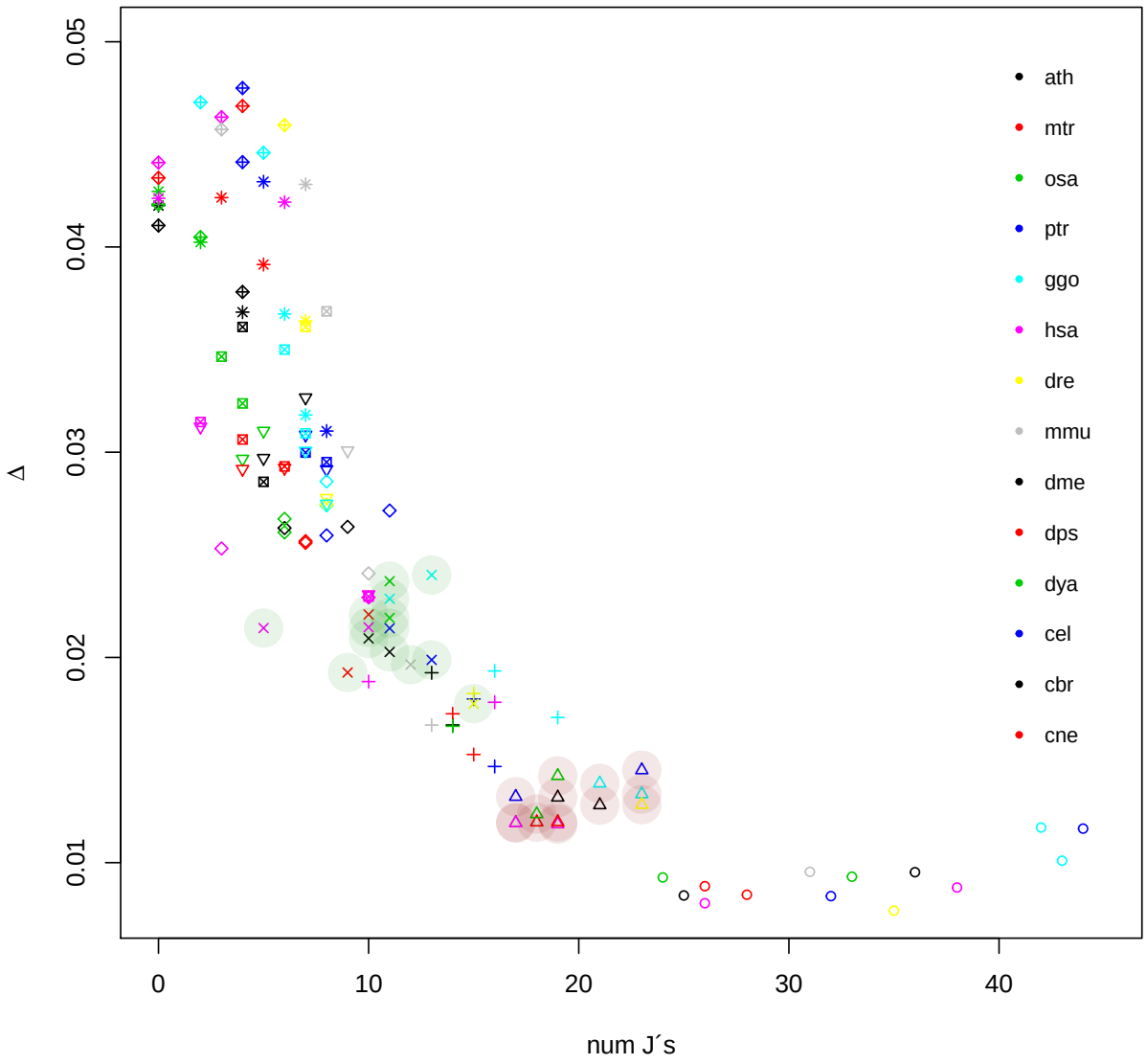
S1 Fig. Size and depth of energy wells in the landscape network. Violin plots for the energy distribution of sequences belonging to different basins of attraction. Each violin plot shows the lowest and highest energies within the corresponding basin and summarizes the overall energy distribution. The first 15 boxplots depict the energy distribution of the basins of the 15 attractor sequences presenting a GT di-nucleotide at the first two intronic positions. The two last boxplots correspond to the energy distribution of non-GT attractor sequences´s basins, and sequences that did not belong to the giant component of the graph respectively. Red dots highlight the attractor (i.e. minimal) energy of each set. The corresponding sequence is included as an x-axis label. Top label show the number of sequences in each basin.

In Sup. Fig S1 Fig we included a boxplot graph displaying the energy distribution of the sequences belonging to different (but maybe overlapping) basins of attraction for the complete set of human donor sequences.

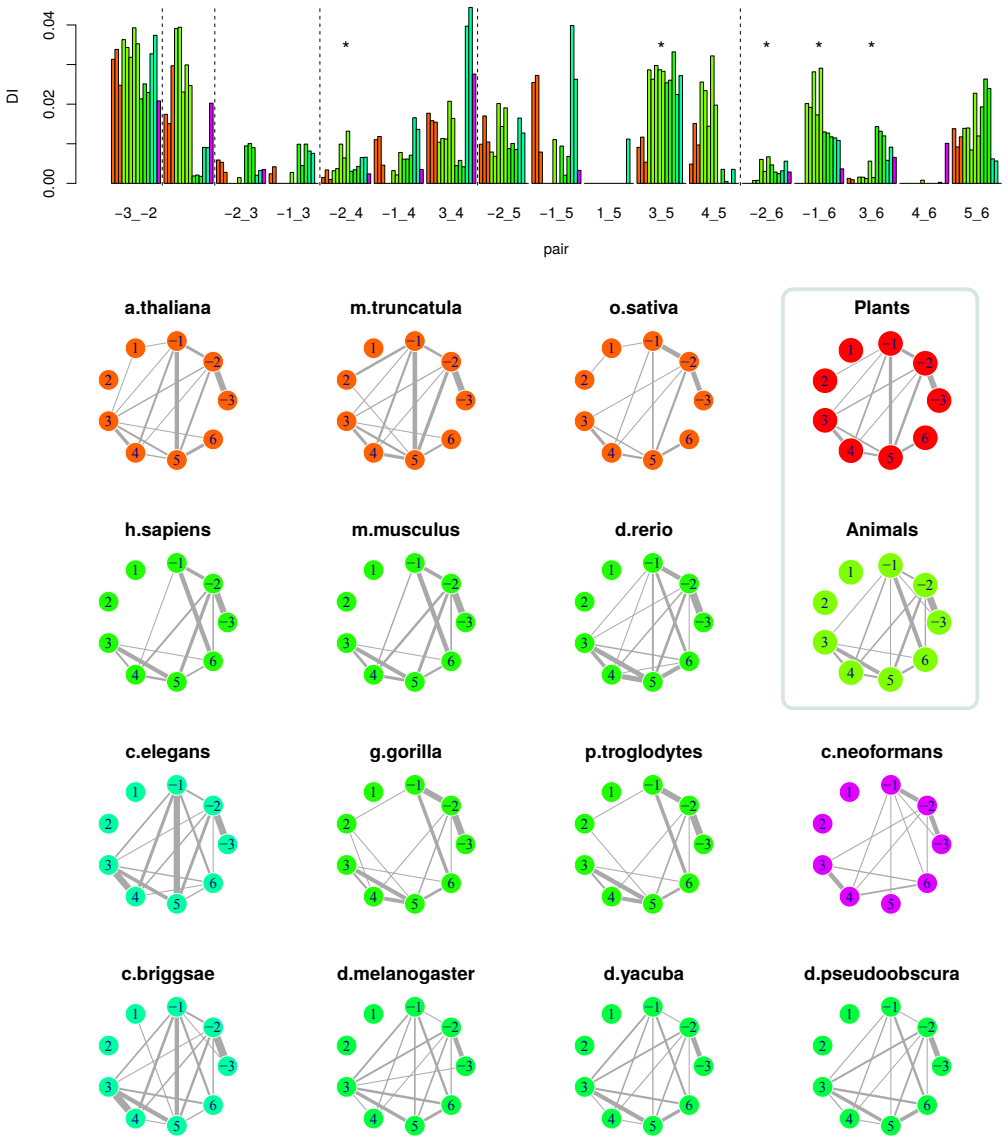
To further characterize the energy landscape of donor sequences, we analized the $\gamma = 0.025$ model for human donor sequences considering a network graph approach. Representing individual sequences with nodes, we connected two of them if they were one mutation away (i.e. Hamming distance of 1). The edge started at the node with higher energy and ended in the lower energy one. In this directed network, we found 243 local-minima sequences associated to nodes with no outwards edges (i.e. null out-degree). For each one of these attractor-nodes we could estimate the in-ward connected component, i.e. the set of nodes that could reach a given local minimum-energy node in a finite number of steps. These sets of nodes were a proxy of the extent of the basin of attraction of a given attractor sequence. In Sup. Fig S1 Fig we included a boxplot graph displaying the energy distribution of the sequences belonging to different (but maybe overlapping) basins of attraction for the complete set of human donor sequences.

The first boxplot in Fig S1 Fig corresponds to the huge main basis of attraction that contained 9593 sequences. This number represented 80% of the complete set of 5' exon-intron boundaries of the *H.sa* genome, and 96% of the entire set of GT 5'ss. Noticeably, the associated attractor state sequence $\vec{S}^* = \{C, A, G, G, T, A, A, G, T\}$ was the perfect matching sequence of U1 smRNA and presented an energy value $E_d(\vec{S}^*) = 3.5$. It can be seen from the figure that the basins of attraction of the rest of the local minima presented much higher energy values in much shallower energy wells. Among these secondary local-minima, there were just 14 attractors presenting GT as first intron nucleotides.

These results suggested that in terms of the model data-driven energy our system presented a rather well-defined global minimum state laying inside a wide energy well.



S2 Fig. Model convergence. Maximum absolute deviation between observed and estimated two site probabilities, $\Delta = max[abs(f_{ij} - P_{ij})]$, as a function of the number of the non-zero coupling constants identified at different regularization levels (different symbols, $\gamma \in [0.01, 0.05]$). Results for different analyzed organisms are depicted with different colors (see legend). Shaded green and red symbols highlight $\gamma = 0.025$ and $\gamma = 0.015$ models respectively.



S3 Fig. Direct Information. Upper panel: Direct information estimates for non-trivial coupling constants found for $\gamma = 0.015$ models. Plants (ath, mtr, osa), vertebrates (ptr, ggo, hsa, dre,mmu), drosophila (dme, dps, dya), nematodes (cel, cbr) and fungal (cne) organism were depicted with red, green, cyan, blue and violet bars respectively. Stars depict site pairs -2:+4. +3:+5, -2:+6, -1:+6,+3:+6, displaying statistically significant differences between plant and animals estimates (Wilcoxon test pvalue < 0.05). Bottom panels: graphical representation of direct information patterns for each analyzed organism. Results obtained for aggregated donor sequences for plants and animals were framed with a gray line. Width edges are proportional to DI values.

organism	code	assembly	5'ss	unique 5'ss	unique GT 5'ss
cryptococcus neoformans	cne	Cryp_neof_125_91_V1	36046	2637	2435
arabidospis thaliana	ath	TAIR10	136036	11286	6803
medicago truncatula	mtr	MedtrA17_4.0	156335	11396	6964
oryza sativa	osa	IRGSP-1.0	130702	12688	8418
pan troglodytes	ptr	Pan_tro_3.0	227807	14903	7919
gorilla gorilla	ggo	gorGor4	212738	14620	7697
homo sapiens	hsa	GRCh38.13	502197	12129	6206
danio rerio	dre	GRCz11.104	276776	11723	6720
mus musculus	mmu	GRCm39.104	391997	9026	5369
caenorhabditis elegans	cel	WBcel235	127661	7625	5351
caenorhabditis briggsae	cbr	CB4	107589	8059	5038
drosophila melanogaster	dme	BDGP6.32	63121	4023	3780
drosophila pseudoobscura	dps	Dpse_3.0	31307	3341	2836
drosophila yacuba	dya	Dyak_caf1	55613	4054	3412

S4 Table. Analyzed genomes. For each analyzed organism we reported, the assembly code, the total number of 5' exon-intron boundaries, the number of different 5'ss sequences and the number of different 5'ss sequences presenting *GT* as the first two intronic bases.

	ath	mtr	osa	ptr	ggo	hsa	dre	mmu	dme	dps	dya	cel	cbr	cne
-3:-2	0.0313	0.0338	0.0248	0.0363	0.0343	0.0318	0.0393	0.0353	0.0213	0.0251	0.0230	0.0327	0.0374	0.0208
-2:-1	0.0174	0.0151	0.0297	0.0391	0.0394	0.0231	0.0299	0.0247	0.0019	0.0021	0.0018	0.0091	0.0090	0.0203
-2:+3	0.0059	0.0053	0.0028	0.0000	0.0000	0.0000	0.0014	0.0000	0.0095	0.0100	0.0090	0.0021	0.0032	0.0035
-1:+3	0.0024	0.0042	0.0000	0.0000	-0.0000	0.0000	0.0028	0.0000	0.0099	0.0045	0.0099	0.0082	0.0076	0.0000
-2:+4	0.0015	0.0034	0.0010	0.0032	0.0037	0.0099	0.0064	0.0132	0.0030	0.0034	0.0043	0.0065	0.0066	0.0024
-1:+4	0.0110	0.0118	0.0046	0.0000	-0.0000	0.0032	0.0021	0.0078	0.0061	0.0061	0.0071	0.0166	0.0136	0.0035
+3:+4	0.0177	0.0158	0.0155	0.0104	0.0113	0.0112	0.0207	0.0164	0.0045	0.0058	0.0042	0.0397	0.0444	0.0276
-2:+5	0.0099	0.0170	0.0104	0.0079	0.0068	0.0202	0.0144	0.0191	0.0088	0.0101	0.0086	0.0165	0.0127	-0.0000
-1:+5	0.0255	0.0273	0.0079	0.0000	-0.0000	0.0000	0.0110	0.0000	0.0094	0.0021	0.0068	0.0398	0.0263	0.0033
+1:+5	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	-0.0000	0.0000	-0.0000	-0.0000	-0.0000	0.0112	0.0000
+3:+5	0.0091	0.0117	0.0053	0.0287	0.0263	0.0298	0.0287	0.0283	0.0254	0.0261	0.0332	0.0225	0.0272	0.0000
+4:+5	0.0049	0.0151	0.0097	0.0256	0.0234	0.0144	0.0322	0.0198	0.0000	0.0035	0.0004	-0.0000	0.0035	-0.0000
-2:+6	-0.0000	0.0000	0.0000	0.0007	0.0008	0.0060	0.0030	0.0067	0.0047	0.0028	0.0025	0.0032	0.0056	0.0029
-1:+6	-0.0000	-0.0000	0.0000	0.0201	0.0192	0.0282	0.0173	0.0291	0.0130	0.0127	0.0118	0.0115	0.0108	0.0037
+3:+6	0.0013	0.0009	0.0000	0.0016	0.0016	0.0012	0.0056	0.0015	0.0143	0.0132	0.0120	0.0058	0.0092	0.0066
+4:+6	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0008	0.0000	0.0000	0.0000	-0.0000	0.0003	-0.0000	0.0101
+5:+6	0.0138	0.0092	0.0117	0.0139	0.0140	0.0084	0.0228	0.0120	0.0193	0.0264	0.0240	0.0062	0.0056	-0.0000

S5 Table. Direct Information. Direct information for two-site interactions. Grayed rows correspond to site-pairs displaying statistically significant differences (Wilcoxon test pvalue < 0.05) between plant and animal estimates.

Species	IC-EC	IC-ENC	INC-EC	INC-ENC	IC-IC	IC-INC	INC-INC	EC-EC	ENC-EC	ENC-ENC
cne	-0.41	0	0	0	-0.14	-0.01	0	0.88	0.13	-0.13
ath	-0.39	0	0.05	0	0.37	0	0	0.83	0	-0.11
mtr	-0.3	0	0	0	0.1	0	0	0.95	0	0
osa	-0.24	0	0.01	0	0.23	0.05	-0.04	0.94	0	0
ptr	-0.19	0	0	0	0.65	-0.01	0	0.73	-0.11	-0.03
ggo	-0.18	0	0	0	0.59	-0.01	0	0.78	-0.1	0
hsa	-0.48	0	0	0	0.14	-0.01	0	0.86	0	0
dre	-0.45	0	0	0	0.23	0	-0.02	0.86	0	0
mmu	-0.51	0	0	0	0.08	0	0	0.86	-0.06	0
cle	-0.9	0	0.05	0	-0.21	-0.01	0	-0.35	0	-0.13
cbr	-0.77	0	0.06	0	0.38	0	0	-0.5	0	-0.14
dme	-0.89	0	0	0	-0.23	-0.04	0	0.38	0	-0.08
dps	-0.66	0	0	0	0.66	-0.05	0	0.36	0	-0.02
dya	-0.77	0	0	0	0.51	-0.04	0	0.37	0	-0.07

S6 Table. Conserved patterns ($\gamma = 0.015$). Mean interactions between different type of sites are shown for different organisms, for a regularization γ value of 0.15. EC, ENC, IC, and INC stand for exonic-consensus, exonic-non-consensus, intronic-consensus and intronic-non-consensus respectively.

Acknowledgments

The authors would like to thank Ezequiel Petrillo and Anabella Srebrow for helpful discussions.

Funding

This work has been supported by grants from Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT). AC also acknowledges support from University of Buenos Aires (grant 20020170100356BA). AC and MY are members of Carrera de Investigador of Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

References

1. G. Ast. How did alternative splicing evolve? *Nature Reviews Genetics*, 5(10):773–782, 2004.

2. W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13):4786–4791, 2012.

3. W. Bialek and R. Ranganathan. Rediscovering the power of pairwise interactions. *arXiv preprint*, (q-bio.QM):1–8, 2007.

4. E. Bouveret, G. Rigaut, A. Shevchenko, M. Wilm, and S. Bertrand. A Sm-like protein complex that participates in mRNA degradation. *The EMBO Journal*, 19(7):1661–1671, 2000.

5. I. Carmel, S. Tal, I. Vig, and G. Ast. Comparative analysis detects dependencies among the 5' splice-site positions. *Rna*, 10(5):828–840, 2004.

6. R. F. Carvalho, D. Szakonyi, C. G. Simpson, I. C. Barbosa, J. W. Brown, E. Baena-González, and P. Duque. The arabidopsis SR45 splicing factor, a negative regulator of sugar signaling, modulates SNF1-related protein kinase 1 stability. *Plant Cell*, 28(8):1910–1925, aug 2016.

7. R. Catalá, C. Carrasco-López, C. Perea-Resa, T. Hernández-Verdeja, and J. Salinas. Emerging Roles of LSM Complexes in Posttranscriptional Regulation of Plant Response to Abiotic Stress. *Frontiers in Plant Science*, 10(167), 2019.

8. S. Chamala, G. Feng, C. Chavarro, and W. B. Barbazuk. Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Frontiers in Bioengineering and Biotechnology*, 3(MAR):33, mar 2015.

9. S. Chaudhary, W. Khokhar, I. Jabre, A. S. Reddy, L. J. Byrne, C. M. Wilson, and N. H. Syed. Alternative splicing and protein diversity: Plants versus animals. *Frontiers in Plant Science*, 10(June):1–14, 2019.

10. W. Chen and M. J. Moore. Spliceosomes, mar 2015.

11. E. Daguenet, G. Dujardin, and J. Valcárcel. The pathogenicity of splicing defects: mechanistic insights into pre mRNA processing inform novel therapeutic approaches. *EMBO reports*, 16(12):1640–1655, 2015.

12. A. De Martino and D. De Martino. An introduction to the maximum entropy approach and its application to inference problems in biology, apr 2018.

13. S. Denisov, G. Bazykin, A. Favorov, A. Mironov, and M. Gelfand. Correlated evolution of nucleotide positions within splice sites in mammals. *PLoS ONE*, 10(12):1–24, 2015.

14. G. Dujardin, C. Lafaille, M. de la Mata, L. E. Marasco, M. J. Muñoz, C. Le Jossic-Corcos, L. Corcos, and A. R. Kornblihtt. How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Molecular Cell*, 54(4):683–690, may 2014.

15. R. Espada, R. G. Parra, T. Mora, A. M. Walczak, and D. U. Ferreiro. Inferring repeat-protein energetics from evolutionary information. *PLoS Computational Biology*, 13(6):1–16, 2017.

16. M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt. How pairwise coevolutionary models capture the collective residue variability in proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, 2018.

17. N. Fong, H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu, and D. L. Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes and Development*, 28(23):2663–2676, dec 2014.

18. E. Ganmor, R. Segev, and E. Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9679–9684, jun 2011.

19. Y. Guo, N. Suzuki, L. Ma, R. Catalá, J. Salinas, C. Perea-Resa, T. Hernández-Verdeja, and C. Carrasco-López. Emerging Roles of LSM Complexes in Posttranscriptional Regulation of Plant Response to Abiotic Stress. *Frontiers in Plant Science*, 2019.

20. H. Iwata and O. Gotoh. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics*, 12(1):45, 2011.

21. J. Jia, Y. Long, H. Zhang, Z. Li, Z. Liu, Y. Zhao, D. Lu, X. Jin, X. Deng, R. Xia, X. Cao, and J. Zhai. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nature Plants*, 6(7):780–788, jul 2020.

22. S. Kandels-Lewis and B. Séraphin. Role of U6 snRNA in 5' splice site selection. *Science*, 262(5142):2035–2039, 1993.

23. E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1):D807–D811, 2019.

24. Z. L, H. J, Z. Y, W. R, L. G, Y. P, Y. C, and S. Y. Crystal structures of the Lsm complex bound to the 3' end sequence of U6 small nuclear RNA. *Nature*, 506(7486):116–120, 2014.

25. T. Laloum, G. Martín, and P. Duque. Alternative Splicing Control of Abiotic Stress Responses. *Trends in Plant Science*, 23(2):140–150, 2018.

26. T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 103(50):19033–19038, dec 2006.

27. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. Technical report, 2011.

28. A. E. Mayes, L. Verdone, P. Legrain, and J. D. Beggs. Characterization of Sm-like proteins in yeast and their association with U6 snRNA. *The EMBO Journal*, 18(15):4321–4331, aug 1999.

29. J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, dec 2012.

30. E. J. Montemayor, J. M. Virta, S. M. Hayes, Y. Nomura, D. A. Brow, and S. E. Butcher. Molecular basis for the distinct cellular functions of the Lsm1-7 and Lsm2-8 complexes. *RNA*, 2020.

31. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–E1301, dec 2011.

32. D. C. Moyer, G. E. Larue, C. E. Hershberger, S. W. Roy, and R. A. Padgett. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Research*, 48(13):7066–7078, jul 2020.

33. T. W. Nilsen. The spliceosome: The most complex macromolecular machine in the cell?, dec 2003.

34. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, dec 2008.

35. A. S. Reddy, Y. Marquez, M. Kalyna, A. Barta, C. W. Reddy, A. S. N, Y. Marquez, M. Kalyna, and A. Barta. Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell*, 25:3657–3683, 2013.

36. F. Remacle, N. Kravchenko-Balasha, A. Levitzki, and R. D. Levine. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22):10324–10329, jun 2010.

37. X. Roca, A. R. Krainer, and I. C. Eperon. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes and Development*, 27(2):129–144, 2013.

38. I. B. Rogozin, L. Carmel, M. Csuros, and E. V. Koonin. Origin and evolution of spliceosomal introns. *Biology Direct*, 7(1):1, 2012.

39. Y. Roudi, E. Aurell, and J. a. Hertz. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3(November):22, 2009.

40. T. S, H. W, M. AE, L. P, B. JD, and P. R. Yeast Sm-like proteins function in mRNA decapping and decay. *Nature*, 404(6777):515–518, mar 2000.

41. K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang, Y. Takeshima, M. Matsuo, G. Sobue, and K. Ohno. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5??? splice sites. *Nucleic Acids Research*, 35(18):5995–6003, sep 2007.

42. M. Santolini, T. Mora, and V. Hakim. A general pairwise interaction model provides an accurate description of in Vivo transcription factor binding sites. *PLoS ONE*, 9(6):1–14, 2014.

43. H. Sawa and J. Abelson. Evidence for a base-pairing interaction between U6 small nuclear RNA and the 5' splice site during the splicing reaction in yeast. *Biochemistry*, 89:11269–11273, 1992.

44. H. Sawa and Y. Shimura. Association of U6 snRNA with the 5'-splice site region of pre-mRNA in the spliceosome. *Genes and Development*, 6(2):244–254, 1992.

45. E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–12, apr 2006.

46. S. H. Schwartz, J. Silva, D. Burstein, T. Pupko, E. Eyras, and G. Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, 18(1):88–103, 2008.

47. M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold, sep 2005.

48. R. M. Stephens and T. D. Schneider. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology*, 228(4):1124–1136, 1992.

49. A. V. Sverdlov, I. B. Rogozin, V. N. Babenko, and E. V. Koonin. Evidence of Splice Signal Migration from Exon to Intron during Intron Evolution. *Current Biology*, 13:2170–2174, 2003.

50. T. A. Thanaraj and A. J. Robinson. Prediction of exact boundaries of exons. *Briefings in bioinformatics*, 1(4):343–356, 2000.

51. G. Tkačik, O. Marre, T. Mora, D. Amodei, M. J. Berry, and W. Bialek. The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(3):P03011, mar 2013.

52. S. Veretnik, C. Wills, P. Youkharibache, R. E. Valas, and P. E. Bourne. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Computational Biology*, 5(3), 2009.

53. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008.

54. M. Weigt, R. a. White, H. Szurmant, J. a. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):67–72, 2009.