Published in final edited form as:

Bioessays. 2013 June ; 35(6): 561–570. doi:10.1002/bies.201200127.

A simple model to explain evolutionary trends of eukaryotic gene architecture and expression:

How competition between splicing and cleavage/polyadenylation factors may affect gene expression and splice-site recognition in eukaryotes

Francesco Catania^{1),*} and Michael Lynch²⁾

¹⁾Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

²⁾Department of Biology, Indiana University, Bloomington, IN, USA

Abstract

Enormous phylogenetic variation exists in the number and sizes of introns in protein-coding genes. Although some consideration has been given to the underlying role of the population-genetic environment in defining such patterns, the influence of the intracellular environment remains virtually unexplored. Drawing from observations on interactions between co-transcriptional processes involved in splicing and mRNA 3'-end formation, a mechanistic model is proposed for splice-site recognition that challenges the commonly accepted intron- and exon-definition models. Under the suggested model, splicing factors that outcompete 3'-end processing factors for access to intronic binding sites concurrently favor the recruitment of 3'-end processing factors at the pre-mRNA tail. This hypothesis sheds new light on observations such as the intron-mediated enhancement of gene expression and the negative correlation between intron length and levels of gene expression.

Keywords

cleavage and polyadenylation; exon definition; gene expression; intron definition; splicing

Introduction

The instructions for the synthesis of proteins are stored within genes in sequences called exons. In eukaryotes and in some viruses, genes often contain additional DNA sequences, called spliceosomal introns, which neither inform nor regulate the assemblage of amino acid chains. Introns are removed from the precursor mRNAs (or pre-mRNAs) through an RNA-splicing process carried out by the spliceosome, a complex molecular machine comprised of proteins and small nuclear RNAs. The spliceosome is guided by critical signals located at the intron termini (the 5' or donor splice site and the 3' or acceptor splice site), and in the intron body (an adenine nucleotide called the *branch site* and a DNA region rich in C and/or U known as the *polypyrimidine tract*). The recognition of and binding to the 5' splice site by

*Corresponding author: Francesco Catania, francesco.catania@uni-muenster.de.

one of the major spliceosomal components, the small nuclear ribonucleoprotein U1 (or U1 snRNP), is typically necessary to trigger the assembly of the spliceosome [1].

RNA splicing is just one of several mRNA-associated processes. 5'-end mRNA capping, mRNA editing, mRNA cleavage and polyadenylation, nuclear export, mRNA surveillance, and mRNA degradation are also associated with the transcription of protein-coding genes. Together with splicing, these processes form a dynamic network of interactions whose structure is currently being elucidated [2], and whose influence on the evolution of the exon-intron gene structure remains virtually unexplored [3].

Here we offer a concise review of the elementary steps in the molecular biology of RNA splicing and its interplay with other mRNA-associated processes, and we present some reasoning on how this interplay could affect the evolution of the intron-exon structure of eukaryotic genes. More specifically, we put forward the hypothesis that gene expression and architecture in eukaryotes are influenced by antagonistic interactions between two major players: (i) the factors that regulate splicing (herein denoted as SFs, for splicing factors), and (ii) the factors that mediate the formation of the mRNA tail or 3' end (herein denoted as CPFs, for cleavage/polyadenylation factors).

The central idea of the proposed hypothesis is that SFs and CPFs compete for access to overlapping or neighboring signal sequences along the transcription unit, particularly within introns and 3' UTRs. This antagonistic relationship likely involves uridine-rich sequences [4, 5] and produces distinct effects. Let us consider, for example, the effects of competition for signal sequences within introns: when SFs efficiently access intronic binding sites, they prevent (or sterically inhibit) the local recruitment of CPFs, thereby enhancing the latter's engagement at the pre-mRNA tail. In contrast, when the engagement of SFs to intronic sequences is inefficient (e.g. due to weak splice signals), CPFs may access introns and encourage a fraction of transcripts to undergo non-canonical splicing events or to remain unspliced (these events are known as alternative splicing). Intron-bound CPFs may remain inoperative (e.g. because of the proximity of U1 snRNP, see below). Alternatively, when key cleavage/polyadenylation signal sequences exist nearby, intron-bound CPFs may promote premature mRNA 3'-end formation (a process known as alternative polyadenylation; Fig. 1).

Several determinants influence the outcome of the competition between splicing factors and cleavage/polyadenylation factors

In the proposed model, the recruitment of competing SFs and CPFs to pre-mRNA binding sites is facilitated by a number of determinants that are able to influence the relative local concentration (or molar ratio) of these two sets of factors. Some of these determinants are described below (see also Fig. 2 and Box 1).

Box 1**Physical interaction of splicing factors and cleavage/polyadenylation factors with the RNA polymerase II**

A number of CPFs and SFs are known to be physically associated [66, 67, 72]. They interact with the phosphorylated carboxy terminal domain of the RNA polymerase II and they may remain attached to the elongation complex throughout transcription [73]. The interaction of CPFs with the carboxy terminal domain may depend on the state of phosphorylation of the Ser 2 and Ser 5 residues within the domain's multiple heptad repeats. In yeast, for example, the ratio of Ser 2 to Ser 5 phosphorylation increases as the RNA polymerase II travels towards the gene 3' end [74]. In addition, phosphorylation of Ser 2 (but not Ser 5) enhances the binding of CPFs [75]. Thus, changes in the ratio of Ser 2 to Ser 5 phosphorylation may contribute to the progressive recruitment of CPFs to the carboxy terminal domain.

At the two ends of the transcription unit, the Cap-Binding Complex and the 3'-end termination signals presumably influence, in opposite ways, the local molar ratio of SFs to CPFs. The Cap-Binding Complex, which binds to 5'-capped RNA polymerase II transcripts, is known to enhance the recruitment of SFs at the mRNA 5'-end [6]. Thus, by increasing the local molar ratio between SFs and CPFs, the Cap-Binding Complex should facilitate splicing in such locations [7, 8]. In contrast, the termination signals that populate mRNA 3'-ends should enhance the recruitment of CPFs, thereby decreasing the local molar ratio of SFs to CPFs and facilitating pre-mRNA 3'-end processing.

Within the transcript unit, strong splicing signals, as well as the proximity of an optimally base-paired U1 snRNP to CPF binding sites (Box 2), are expected to favor splicing. A pronounced exon-intron differential GC content may also favor splicing by guiding the recruitment of serine/arginine-rich proteins that bind AG-rich or AC-rich sequence elements [9, 10]. Serine/arginine-rich proteins typically assist the recruitment of spliceosomal components to the proximal splice sites when they bind within exons, whereas they act as splicing inhibitors when they bind within introns [11–13]. Thus, the combined presence of GC-rich exons and GC-poorer flanking introns should facilitate the specific binding of serine/arginine-rich proteins to exonic target sites, thereby enhancing splicing.

Box 2**The distance-dependent inhibitory effect of U1 snRNP**

Studies of viruses, plants, vertebrates, and invertebrates show that the 5' splice site has an inhibitory effect on cleavage/polyadenylation. Such an effect is: (i) distance-dependent; (ii) mediated by a base-paired U1 snRNP; and (iii) manifested both within introns and nearby transcription termination signals at the mRNA tail [5, 16, 58, 70, 76–79].

The U1 snRNP can inhibit the downstream function of RNA-bound CPFs [20] and/or their physical binding [80]. Also, an optimal binding of U1 snRNP with the donor splice site may facilitate the inhibitory effect on CPFs [81]. The U1 snRNP-mediated influence

appears to be particularly (but not exclusively) effective when the spacing between the optimal donor site and the CPF binding sites is $< \sim 1,000$ nt [16, 20, 70, 76, 77, 82]. Finally, the inhibitory effect of U1 snRNP may be bidirectional, i.e. it influences sequences that are located both downstream and upstream of the 5' splice site. For intronic sequences that are upstream of the 5' splice site, the strength of the U1 snRNP inhibitory effect would be mediated by the length of the intervening exon, as hinted by previous experimental work [45].

While the list of determinants presented here is probably not exhaustive, it is worth noting that determinants need not be RNA elements; they may also involve the transcribed DNA, as exemplified by recent studies which demonstrate that histone modifications facilitate the recruitment of SFs [14, 15].

Competing splicing factors and cleavage/polyadenylation factors promote coupled processes

Although an antagonistic relationship between SFs and CPFs is widely reported in the literature [5, 16–20], the processes of splicing and cleavage/polyadenylation are known to be coupled [21]. Together, these observations raise the obvious question: how can competing factors promote cooperative processes?

A real-life example may help reconcile these seemingly conflicting observations. Imagine a situation in which an individual performs two tasks either at the same time or sequentially. If she multitasks, each task is performed inefficiently because of competition for common energy resources. In contrast, if she performs the two tasks sequentially, there is no competition (or simultaneous engagement with common resources) and the efficiency of each task is enhanced. When considered like this, the two tasks appear to be coupled.

When applied to the intracellular environment, the proposed example may be read as follows: competition between SFs and CPFs for access to pre-mRNA binding sites hampers both splicing and mRNA 3'-end formation. In contrast, reduced antagonism between SFs and CPFs – which may occur when determinants inhibit the access of CPFs to intronic sites or of SFs to 3'-end termination signals – enhances the processes of splicing and mRNA 3'-end formation.

Under this scenario, the introduction of efficiently spliced introns into intronless transcripts is expected to reduce competition at the pre-mRNA 3'-end, and ultimately enhance transcriptional yield. Remarkably, this prediction corresponds with (and may shed light on the underlying mechanism of) a commonly observed phenomenon called “intron-mediated enhancement of gene expression” [22, 23].

Why are genes with small introns typically more highly expressed than genes with large introns?

The scenario proposed above may yield insight into the longstanding question of why genes with small introns are typically more highly expressed compared to genes with large introns [24–27] (see also Box 3). At least three explanations have been invoked to answer this question: (i) selection acts to minimize the cost of transcription [26]; (ii) selection acts against large introns in active chromosomal compartments [28]; and (iii) selection favors less complex regulation and architecture of housekeeping genes [29].

Box 3

Exon size, rather than intron size, may contribute to gene expression in plants

The inverse relationship between intron length and gene expression observed in animals does not appear to hold for plants; in plants, highly expressed genes contain larger introns compared to lowly expressed genes [83]. These seemingly contrasting observations can be reconciled with our model. To begin, the average size of *large* introns for highly expressed genes in rice and Arabidopsis is only 416 and 164 nucleotides (nt; as opposed to 359 and 140 nt for lowly expressed genes) [83]. This intron size is considerably smaller than that of large animal introns –which can extend for several kbs –and, importantly, is smaller than the observed distance covered by the inhibitory action of U1 snRNP on the access of CPFs to downstream intronic target sites (i.e. ~500 nt). This latter observation implies that the intron size in rice and Arabidopsis may not be a major determinant of gene expression.

On the other hand, in rice and Arabidopsis, lowly expressed genes have larger exons (479 and 471 nt on average) compared to highly expressed genes (372 and 329 nt on average, respectively) [83]. Under our competition model, large exons favor the access of CPFs to upstream intronic target sites – especially if the 5' splice site of the upstream intron is weak (Fig. 5) – disfavoring the recruitment of these factors at the pre-mRNA 3' tail. Thus, our model predicts that exon size, more than intron size, is relevant to gene expression in plants.

While we acknowledge the plausibility of each of these hypotheses, we propose that the negative correlation between expression level and intron size is a direct byproduct of molecular interactions. Specifically, we suggest that SFs are able to outcompete CPFs for access to intronic binding sites more efficiently within small introns than within large introns. Under this scenario, CPFs that are effectively outcompeted for binding intronic sites in short-intron-containing transcripts are more efficiently engaged to the pre-mRNA tail and enhance transcription yield.

Why is CPF recruitment to intronic binding sites less likely in short introns than in large introns? We suggest that the distance-dependent inhibitory effect of U1 snRNP on CPFs provides a feasible explanation (Box 2). While it is difficult to accurately quantify how rapidly the inhibitory effect of U1 snRNP decays with distance, it seems reasonable to

propose that, all else being equal, this inhibitory effect is reduced in naturally large introns or artificially expanded small introns. In these cases, the access of CPFs to intronic binding sites would be facilitated and alternative splicing may occur [30, 31]. This event would effectively promote less efficient recruitment of CPFs at the tail of the transcript and result in low or moderate transcription yield (Fig. 3).

Trade-offs between the determinants of competition influence the antagonistic relationships between splicing factors and cleavage/polyadenylation factors

Under our model, if large introns contain a weak 5' splice site and key termination signals then CPFs are more likely to access intronic binding sites and promote premature mRNA 3'-end formation. This expectation is consistent with events observed during alternative polyadenylation in the introns of humans (3,344 genes involved [32]) and Arabidopsis (2,100 genes involved [33]), which are significantly larger (medians: 3,236 nt vs. 1,552 nt in humans; 270 nt vs. 99 nt in Arabidopsis) and have weaker 5' splice sites (but comparably strong 3' splice sites) compared to poly(A)-free introns. Although genes with large introns undergo accurate splicing less often than genes with short introns [34], large introns can, nonetheless, be correctly spliced. This suggests that other variables must come into play to guarantee correct splicing besides the proximity of an optimally base-paired U1 snRNP to CPF binding sites.

Under our model, large introns that are accurately spliced should exhibit and/or be in proximity of signals that increase the local molar ratio between SFs and CPFs (Fig. 2). In remarkable agreement with these predictions, splice-site strength has been found to scale positively with intron size across numerous species [35–38]. In addition, large introns are typically located in genic regions where we expect the recruitment of SFs to be enhanced, e.g. in proximity of the Cap-Binding Complex and beside GC-richer exons. Indeed, 5' UTR introns are typically >twofold larger, on average, than CDS or 3' UTR introns [39], whereas first CDS introns tend to be ~40% larger than other CDS introns [40]. Also, in several eukaryotes the GC-content differential between exons and their flanking introns is more pronounced for large introns than for short introns [41].

The intracellular and population-genetic environments are linked

Thus far we have proposed that the antagonistic relationship between SFs and CPFs may help explain a number of trends in gene expression and architecture across eukaryotes. Under our model, this antagonistic relationship is mediated by trade-offs between determinants of competition (e.g. splice-site strength, exon-intron differential GC content), whose interactions determine a dynamic equilibrium. Deviations from this equilibrium (e.g. due to weak selective pressures) are expected to perturb gene structure, and generate non-canonical or alternative splicing isoforms or intronic polyadenylation events, of which a large number are presumably non-functional and thus subject to purifying selection. This latter scenario leads to an intriguing prediction: because the efficiency of natural selection decreases with increasing organism size [42], the fraction of intronic polyadenylation and

alternative splicing events in eukaryotes should gradually increase from unicellular to multicellular species. While there is currently not enough information available to explore trends for intronic polyadenylation, estimates for the prevalence of alternative splicing across multiple eukaryotes appear to support our model prediction (Fig. 4).

How are splice sites recognized?

Two mechanisms have been invoked to explain how the spliceosome selects splice sites: the intron- and the exon-definition models. According to the former, introns are the spliceosome's units of recognition and the spliceosome identifies the 5' and 3' splice sites flanking the intron [43]. Mutation of a splice site at one end of the intron would inhibit splicing and lead to intron retention. In exon-definition [44, 45], 5' and 3' splice sites are recognized in concert across an exon. In exon-definition, the spliceosomal machinery would search for and bind splice sites at the ends of an internal exon, and mutation of a splice site at one end of this exon would lead to exon skipping.

Perhaps the most compelling evidence to support the exon-definition mechanism is that mutational inactivation of the 5' splice site of a downstream intron represses splicing of the upstream intron [46]. In an intron-defined mechanism, the disruption of a splice site would be isolated to the intron that contains the mutation. In contrast, under exon-definition, the mutation of a 5' splice site downstream of an internal exon would inhibit recognition of the exon, producing both the observed suppression of splicing and exon skipping.

Recognition of terminal exons is problematic for the exon-defined mechanism because these exons lack either the 3' splice site or 5' splice site. While the definition of terminal exons is suggested to be mediated by the 5' mRNA capping complex and factors involved in mRNA maturation, respectively [44], how this mediation might occur remains unclear.

Exon-definition is commonly thought to occur in vertebrates where exon size is typically much shorter than intron size, while intron-definition is thought to occur in eukaryotes with short introns. However, exon skipping can also occur (even if infrequently) in species with typically short introns, such as *Arabidopsis* [47]. The opposite is true for intron retention, which is also found in species with large introns such as humans [21]. This raises some doubts as to whether intron (exon) size alone mediates intron (exon)-definition, as previously proposed [30]. In vitro studies suggest that the unit of definition may switch from intron to exon when intron size reaches ~250 nucleotides [31], but this may not always be the case [48]. At least one case of splicing has been reported where neither of the two models completely accounts for the experimental observations [49], and the evolutionary relationship between the two mechanisms of splice-site recognition is uncertain [50]. Furthermore, although particular attention has been given to the putative transition from exon definition to intron definition [51], the mechanisms that would allow the two modes to coexist (as they seem to in various species [52]) and to alternate in genes or species with both short and long introns [30] have not yet been identified.

U1-dependent definition

We propose a novel mechanism for the *modus operandi* of splicing, U1-dependent definition, in which aspects of both an exon and its upstream intron elucidate the selection of splice sites. The mechanism we put forward relies on the distance-dependent inhibitory effect that the U1 snRNP exerts on CPFs as well as on the central role that the U1 snRNP plays in the formation of the splicing commitment complex [1, 53] and in the binding of additional SFs to 3' splice sites [54–56].

In the following sections we illustrate how U1-dependent definition can convincingly explain changes in splice-site recognition that result from the strength of 5' splice sites and gene architecture. It is worth noting here that U1-dependent definition provides a logical explanation for the findings of Sterner et al. [57], a well-designed study that represents a cornerstone for the intron- and exon-definition models.

Internal exons

As we have discussed above, suboptimal splicing conditions (e.g. weak splice signals, large intron size, and poor intron-exon GC-content differential) encourage the access of CPFs to introns; this facilitates the occurrence of alternative splicing events, such as the skipping of the downstream exon. If this downstream exon is sufficiently small, however, and the 5' splice site of the intron located downstream from this exon has optimal base-pairing with U1 snRNP, then it is unlikely that the skipping event will take place. Under these conditions, this U1 snRNP would exert an inhibitory effect on the access of CPFs to the proximal upstream intronic sites, favoring splicing [57, 58] (Fig. 5A). If the exon is small, but the 5' splice site of the downstream intron is suboptimal, then the U1 snRNP binding would be less favored and alternative splicing (or premature mRNA 3'-end formation) is possibly promoted (Fig. 5B). In this view, the size expansion of the downstream exon would tend to repress the distance-dependent inhibitory effect of U1 snRNP on CPFs and favor alternative splicing (or premature mRNA 3'-end formation) despite the presence of an optimal downstream 5' splice site [57] (Fig. 5C). Notably, the expansion of a small exon and/or a mutational deactivation of the 5' splice site downstream of an internal exon would upset the splicing of the upstream intron, unless the latter is short. In such a case, the U1 snRNP bound to the 5' splice site of the upstream intron would have an inhibitory effect on the access of CPFs to the downstream intronic binding sites, thereby favoring splicing [59, 60] (Fig. 5D).

Terminal exons

Berget [44] suggests that the Cap-Binding Complex plays a role in the splicing of the first intron; from the exon perspective, this implies that the Cap-Binding Complex and the first 5' splice site may be recognized in concert. How this would take place is yet to be shown. As for the *definition* of last exons, *in vitro* and *in vivo* observations are consistent with the idea that the ends of these terminal exons – a 3' splice site and a poly(A) site – are recognized in concert. More specifically, it has been reported that while a mutated 3' poly(A) site impairs splicing of the last intron, a mutated terminal 3' splice site or polypyrimidine tract disfavors or inhibits polyadenylation [48, 61–64].

Taking the molecular interactions discussed above into account, we propose that first-intron splicing is assisted by the Cap-Binding Complex-enhanced recruitment of SFs at the pre-mRNA 5'-end. At the other end of the transcript, splicing of last introns may be facilitated by 3'-end termination signals that recruit CPFs and in so doing help SFs to outcompete CPFs within the last intron. Under this scenario, mutated 3'-end transcription termination signals inhibit last intron splicing because they ineffectively recruit CPFs. Similarly, mutated splicing signals of the last introns perturb the process of mRNA maturation because they intensify the antagonism between SFs and CPFs that have overlapping or neighboring binding sites in the last exon (e.g. U-rich sequences located upstream of the cleavage site [65]). These hypothetical dynamics may explain the interdependence between 3'-end processing and splicing observed by earlier studies [66, 67].

Additional layers of complexity may be added to the scenario described above. For example, the close proximity of the last intron to 3'-end termination signals probably disfavors rather than facilitates splicing, in that the termination signals that efficiently recruit CPFs would increase the local molar ratio of CPFs to SFs. Observations in yeast support this scenario. Specifically, Tardiff et al. [68] measured the efficiency with which two SFs, U1 snRNP and U2 snRNP, are recruited to the 5' splice site and the branch site, respectively, of an intron within two-exon constructs. The second exon in these constructs has a variable length (ranging from 350 to 2,300 bp) and contains a 3' UTR segment with functional termination signals. Their results suggest that the levels of recruitment of U2 snRNP are considerably higher in constructs with larger second exons. Also, premature cleavage and polyadenylation is likely to take place in constructs with short second exons [68]. These results are consistent with the idea that the 3'-end termination signals compromise the (co-transcriptional) splicing of introns that are close to the tail of the transcript.

Open questions and limitations of U1-dependent definition

Our model makes several predictions (Box 4) but leaves a number of central questions unresolved, some of which are listed below.

1. What are the nature and the relative locations of the CPF intronic binding sites? While we propose that U-rich sequences are suitable candidates [4, 5], other motifs may be involved. Also, do CPF intronic binding sites preferentially reside in specific portions of the intron (e.g. flanking or overlapping the polypyrimidine tract), or are they distributed throughout the intron [69]?
2. What is the range of the U1 snRNP inhibitory action on the access of CPFs to downstream binding sites? While a recent study shows that this action extends up to ~500–1,000 nt in humans, mouse, and *Drosophila* [70], estimates for other eukaryotic groups await formal investigation.
3. Is the inhibitory effect of U1 snRNP bidirectional as it is assumed above? And if so, is the strength of this effect comparable with that exerted on upstream binding sites?

Box 4**Model predictions**

In addition to those discussed in the text, a number of further verifiable predictions logically arise from our model. Some are listed below:

1. Because under our competition model the recruitment of SFs along the transcription unit is facilitated at the 5'-end more than at the 3'-end, spliceosomal introns should preferentially occur in 5' UTRs compared to 3' UTRs. This prediction is consistent with results obtained for human, mouse, Arabidopsis, and Drosophila [39].
2. Under our model, one may expect that a trade-off exists between splicing-enhancing conditions (e.g. strong splice signals) and conditions that disfavor splicing (e.g. large downstream exon size). Consistent with this idea, large exons are often associated with short upstream introns in humans [84]. Moreover, in human and Drosophila the strength of the 5' splice site of the next downstream intron increases with the length of the upstream intron [85].
3. A number of SFs and CPFs would be expected to act independently of their binding to canonical target signals. Consistent with this idea, the splicing factor U2AF 65 is able to bind to many intronless mRNAs [86]; several splicing factors bind to non-intronic U-rich sequence elements that reside upstream of the cleavage site [65]; components of the cleavage and polyadenylation specificity factor bind U-rich sequences that reside upstream of the poly(A) signal [4]; and nearly half of all the binding sites of the cleavage stimulation factor 64 in human cells reside within introns [5].
4. The inhibited recruitment of SFs at the pre-mRNA tail should favor retention of introns in 3' UTRs. This prediction is consistent with results obtained for humans and Arabidopsis [87, 88].

Unfortunately, it is difficult to disentangle the effects of U1-dependent definition from those of exon- and intron-definition, as these effects coincide in many cases. This is made most clear by the previously discussed observations of Steiner et al. [57], which can be explained equally well under exon- and intron-definition and U1-dependent definition. The development of strategies to examine the validity of U1-dependent definition is under way, and at least two additional observations make the investigation of this theoretical mechanism of definition worthwhile. First, in humans, selective constraints exist on the length of exons and their flanking introns [71]. Second, in Drosophila and humans, the length of the upstream intron is significantly more important than the length of the downstream intron in determining whether or not the encompassed exon will be skipped [31]. These observations hint at the possibility that the spliceosome recognizes an intron and the following exon as a unit. While not expected under the exon- and intron-definition models, this is consistent with U1-dependent definition (Fig. 5).

Conclusions

The ideas outlined above provide a broad and coherent view on how interacting transcriptional processes may affect the evolution of gene expression and architecture in eukaryotes. What we accomplish is the formulation of a coherent molecular scenario wherein the competition for access to pre-mRNA binding sites between splicing factors and cleavage/polyadenylation factors is one of the major driving forces in the selection of exons and transcription termination sites. Such a scenario (i) makes predictions that extend to the selective environment to which distinct eukaryotic lineages are subject [42], and (ii) integrates logically with (and extends) the proposal that the intracellular environment contributes to the physical establishment of spliceosomal introns in eukaryotic genes [3].

We have presented simple and logical connections to frame observations such as intron-mediated enhancement of gene expression, negative correlation between intron size and levels of gene expression, and exon- and intron-definition modes of splice-site recognition in the context of the proposed model. One key influential factor is the U1 snRNP, which interacts with 5' splice sites and exerts a distance-dependent inhibitory effect on cleavage/polyadenylation factors. The proposed U1-dependent definition potentially unifies the intron- and exon-definition models and, if experimentally verified, would imply that the mechanisms underlying splice-site recognition across eukaryotes are more similar than currently thought.

Acknowledgments

We thank J. Schmitz for comments on an earlier version of this manuscript. This work was supported by a Marie Curie International Incoming Fellowship (grant 254202) awarded to F.C., MetaCyte funding from the Lilly Foundation to Indiana University, and the National Science Foundation grant EF-0827411 to M.L.

Abbreviations

CPFs	cleavage/polyadenylation factors
SFs	splicing factors
U1 snRNP	small nuclear ribonucleoprotein U1

References

1. Seraphin B, Rosbash M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell*. 1989; 59:349–58. [PubMed: 2529976]
2. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature*. 2002; 416:499–506. [PubMed: 11932736]
3. Catania F, Lynch M. Where do introns come from? *PLoS Biol*. 2008; 6:e283. [PubMed: 19067485]
4. Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep*. 2012; 1:753–63. [PubMed: 22813749]
5. Yao C, Biesinger J, Wan J, Weng L, et al. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc Natl Acad Sci USA*. 2012; 109:18773–8. [PubMed: 23112178]

6. Lewis JD, Izaurralde E, Jarmolowski A, McGuigan C, et al. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev.* 1996; 10:1683–98. [PubMed: 8682298]
7. Ohno M, Sakamoto H, Shimura Y. Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure. *Proc Natl Acad Sci USA.* 1987; 84:5187–91. [PubMed: 2440046]
8. Inoue K, Ohno M, Sakamoto H, Shimura Y. Effect of the cap structure on pre-mRNA splicing in *Xenopus* oocyte nuclei. *Genes Dev.* 1989; 3:1472–9. [PubMed: 2606355]
9. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science.* 2002; 297:1007–13. [PubMed: 12114529]
10. Schaal TD, Maniatis T. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol.* 1999; 19:1705–19. [PubMed: 10022858]
11. Kanopka A, Muhlemann O, Akusjarvi G. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature.* 1996; 381:535–8. [PubMed: 8632829]
12. Ibrahim el C, Schaal TD, Hertel KJ, Reed R, et al. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci USA.* 2005; 102:5002–7. [PubMed: 15753297]
13. Erkelenz S, Mueller WF, Evans MS, Busch A, et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA.* 2013; 19:96–102. [PubMed: 23175589]
14. Luco RF, Pan Q, Tominaga K, Blencowe BJ, et al. Regulation of alternative splicing by histone modifications. *Science.* 2010; 327:996–1000. [PubMed: 20133523]
15. Pradeepa MM, Sutherland HG, Ule J, Grimes GR, et al. Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.* 2012; 8:e1002717. [PubMed: 22615581]
16. Qiu J, Pintel DJ. Alternative polyadenylation of adeno-associated virus type 5 RNA within an internal intron is governed by the distance between the promoter and the intron and is inhibited by U1 small nuclear RNP binding to the intervening donor. *J Biol Chem.* 2004; 279:14889–98. [PubMed: 14749332]
17. Phillips C, Pachikara N, Gunderson SI. U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. *Mol Cell Biol.* 2004; 24:6162–71. [PubMed: 15226420]
18. Evsyukova I, Bradrick SS, Gregory SG, Garcia-Blanco MA. Cleavage and polyadenylation specificity factor 1 (CPSF1) regulates alternative splicing of interleukin 7 receptor (IL7R) exon 6. *RNA.* 2013; 19:103–15. [PubMed: 23151878]
19. Su Y, Adair R, Davis CN, DiFronzo NL, et al. Convergence of RNA *cis* elements and cellular polyadenylation factors in the regulation of human cytomegalovirus UL37 exon 1 unspliced RNA production. *J Virol.* 2003; 77:12729–41. [PubMed: 14610195]
20. Gunderson SI, Polycarpou-Schwarz M, Mattaj IW. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Mol Cell.* 1998; 1:255–64. [PubMed: 9659922]
21. Wang ET, Sandberg R, Luo S, Khrebtkova I, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 2008; 456:470–6. [PubMed: 18978772]
22. Rose AB, Emami S, Bradnam K, Korf I. Evidence for a DNA-based mechanism of intron-mediated enhancement. *Front Plant Sci.* 2011; 2:98. [PubMed: 22645558]
23. Morello L, Giani S, Troina F, Breviario D. Testing the IMETER on rice introns and other aspects of intron-mediated enhancement of gene expression. *J Exp Bot.* 2011; 62:533–44. [PubMed: 20855457]
24. Seoighe C, Gehring C, Hurst LD. Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet.* 2005; 1:e13. [PubMed: 16110339]
25. Urrutia AO, Hurst LD. The signature of selection mediated by expression on human genes. *Genome Res.* 2003; 13:2260–4. [PubMed: 12975314]

26. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, et al. Selection for short introns in highly expressed genes. *Nat Genet.* 2002; 31:415–8. [PubMed: 12134150]
27. Carmel L, Koonin EV. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome Biol Evol.* 2009; 1:382–90. [PubMed: 20333206]
28. Prachumwat A, DeVincentis L, Palopoli MF. Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics.* 2004; 166:1585–90. [PubMed: 15082572]
29. Vinogradov AE. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 2004; 20:248–53. [PubMed: 15109779]
30. Talerico M, Berget SM. Intron definition in splicing of small *Drosophila* introns. *Mol Cell Biol.* 1994; 14:3434–45. [PubMed: 8164690]
31. Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, et al. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci USA.* 2005; 102:16176–81. [PubMed: 16260721]
32. Tian B, Pan Z, Lee JY. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 2007; 17:156–65. [PubMed: 17210931]
33. Wu X, Liu M, Downie B, Liang C, et al. Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive alternative polyadenylation. *Proc Natl Acad Sci USA.* 2011; 108:12533–8. [PubMed: 21746925]
34. Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 2007; 35:125–31. [PubMed: 17158149]
35. Weir M, Rice M. Ordered partitioning reveals extended splice-site consensus information. *Genome Res.* 2004; 14:67–78. [PubMed: 14707171]
36. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, et al. Introns and splicing elements of five diverse fungi. *Eukaryot Cell.* 2004; 3:1088–100. [PubMed: 15470237]
37. Fahey ME, Higgins DG. Gene expression, intron density, and splice site strength in *Drosophila* and *Caenorhabditis*. *J Mol Evol.* 2007; 65:349–57. [PubMed: 17763878]
38. Dewey CN, Rogozin IB, Koonin EV. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics.* 2006; 7:311. [PubMed: 17156453]
39. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol.* 2006; 23:2392–404. [PubMed: 16980575]
40. Bradnam KR, Korf I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE.* 2008; 3:e3093. [PubMed: 18769727]
41. Amit M, Donyo M, Hollander D, Goren A, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* 2012; 1:543–6. [PubMed: 22832277]
42. Lynch M. The origins of eukaryotic gene structure. *Mol Biol Evol.* 2006; 23:450–68. [PubMed: 16280547]
43. Romfo CM, Alvarez CJ, van Heeckeren WJ, Webb CJ, et al. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol Cell Biol.* 2000; 20:7955–70. [PubMed: 11027266]
44. Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem.* 1995; 270:2411–4. [PubMed: 7852296]
45. Robberson BL, Cote GJ, Berget SM. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol.* 1990; 10:84–94. [PubMed: 2136768]
46. Talerico M, Berget SM. Effect of 5' splice site mutations on splicing of the preceding intron. *Mol Cell Biol.* 1990; 10:6299–305. [PubMed: 2247057]
47. Marquez Y, Brown JW, Simpson C, Barta A, et al. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* 2012; 22:1184–95. [PubMed: 22391557]

48. Rigo F, Martinson HG. Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. *Mol Cell Biol.* 2008; 28:849–62. [PubMed: 17967872]
49. Howe KJ, Kane CM, Ares M Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA.* 2003; 9:993–1006. [PubMed: 12869710]
50. Ram O, Ast G. SR proteins: a foot on the exon before the transition from intron to exon definition. *Trends Genet.* 2007; 23:5–7. [PubMed: 17070958]
51. Sharma S, Kohlstaedt LA, Damianov A, Rio DC, et al. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat Struct Mol Biol.* 2008; 15:183–91. [PubMed: 18193060]
52. McGuire AM, Pearson MD, Neafsey DE, Galagan JE. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol.* 2008; 9:R50. [PubMed: 18321378]
53. Michaud S, Reed R. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev.* 1991; 5:2534–46. [PubMed: 1836445]
54. Kuo HC, Nasim FH, Grabowski PJ. Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. *Science.* 1991; 251:1045–50. [PubMed: 1825520]
55. Izquierdo JM, Majos N, Bonnal S, Martinez C, et al. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell.* 2005; 19:475–84. [PubMed: 16109372]
56. Sharma S, Falick AM, Black DL. Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex. *Mol Cell.* 2005; 19:485–96. [PubMed: 16109373]
57. Sterner DA, Carlo T, Berget SM. Architectural limits on split genes. *Proc Natl Acad Sci USA.* 1996; 93:15081–5. [PubMed: 8986767]
58. Abad X, Vera M, Jung SP, Oswald E, et al. Requirements for gene silencing mediated by U1 snRNA binding to a target sequence. *Nucleic Acids Res.* 2008; 36:2338–52. [PubMed: 18299285]
59. Bell MV, Cowper AE, Lefranc MP, Bell JI, et al. Influence of intron length on alternative splicing of CD44. *Mol Cell Biol.* 1998; 18:5930–41. [PubMed: 9742110]
60. Peterson ML, Perry RP. Regulated production of mu m and mu s mRNA requires linkage of the poly(A) addition sites and is dependent on the length of the mu s-mu m intron. *Proc Natl Acad Sci USA.* 1986; 83:8883–7. [PubMed: 3097638]
61. Liu X, Mertz JE. Polyadenylation site selection cannot occur in vivo after excision of the 3'-terminal intron. *Nucleic Acids Res.* 1993; 21:5256–63. [PubMed: 8255783]
62. Niwa M, Rose SD, Berget SM. In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev.* 1990; 4:1552–9. [PubMed: 1701407]
63. Niwa M, Berget SM. Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes Dev.* 1991; 5:2086–95. [PubMed: 1657710]
64. Cooke C, Hans H, Alwine JC. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. *Mol Cell Biol.* 1999; 19:4971–9. [PubMed: 10373547]
65. Millevoi S, Vagner S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res.* 2010; 38:2757–74. [PubMed: 20044349]
66. Kyburz A, Friedlein A, Langen H, Keller W. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Mol Cell.* 2006; 23:195–205. [PubMed: 16857586]
67. Millevoi S, Loulergue C, Dettwiler S, Karaa SZ, et al. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. *EMBO J.* 2006; 25:4854–64. [PubMed: 17024186]
68. Tardiff DF, Lacadie SA, Rosbash M. A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Mol Cell.* 2006; 24:917–29. [PubMed: 17189193]
69. Rose AB, Elfersi T, Parra G, Korf I. Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *Plant Cell.* 2008; 20:543–51. [PubMed: 18319396]

70. Berg MG, Singh LN, Younis I, Liu Q, et al. U1 snRNP determines mRNA length and regulates isoform expression. *Cell*. 2012; 150:53–64. [PubMed: 22770214]
71. Schwartz S, Gal-Mark N, Kfir N, Oren R, et al. Alu exonization events reveal features required for precise recognition of exons by the splicing machinery. *PLoS Comput Biol*. 2009; 5:e1000300. [PubMed: 19266014]
72. Awasthi S, Alwine JC. Association of polyadenylation cleavage factor I with U1 snRNP. *RNA*. 2003; 9:1400–9. [PubMed: 14561889]
73. Phatnani HP, Greenleaf AL. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev*. 2006; 20:2922–36. [PubMed: 17079683]
74. Komarnitsky P, Cho EJ, Buratowski S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev*. 2000; 14:2452–60. [PubMed: 11018013]
75. Licatalosi DD, Geiger G, Minet M, Schroeder S, et al. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Mol Cell*. 2002; 9:1101–11. [PubMed: 12049745]
76. Ashe MP, Pearson LH, Proudfoot NJ. The HIV-1 5' LTR poly(A) site is inactivated by U1 snRNP interaction with the downstream major splice donor site. *EMBO J*. 1997; 16:5752–63. [PubMed: 9312033]
77. Kaida D, Berg MG, Younis I, Kasim M, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*. 2010; 468:664–8. [PubMed: 20881964]
78. Vagner S, Rueggsegger U, Gunderson SI, Keller W, et al. Position-dependent inhibition of the cleavage step of pre-mRNA 3'-end processing by U1 snRNP. *RNA*. 2000; 6:178–88. [PubMed: 10688357]
79. Luehrsen KR, Walbot V. Intron creation and polyadenylation in maize are directed by AU-rich RNA. *Genes Dev*. 1994; 8:1117–30. [PubMed: 7926791]
80. Niwa M, MacDonald CC, Berget SM. Are vertebrate exons scanned during splice-site selection? *Nature*. 1992; 360:277–80. [PubMed: 1359430]
81. Qiu J, Cheng F, Pintel D. Distance-dependent processing of adeno-associated virus type 5 RNA is controlled by 5' exon definition. *J Virol*. 2007; 81:7974–84. [PubMed: 17507471]
82. Fortes P, Cuevas Y, Guan F, Liu P, et al. Inhibiting expression of specific genes in mammalian cells with 5' end-mutated U1 small nuclear RNAs targeted to terminal exons of pre-mRNA. *Proc Natl Acad Sci USA*. 2003; 100:8264–9. [PubMed: 12826613]
83. Ren XY, Vorst O, Fiers MW, Stiekema WJ, et al. In plants, highly expressed genes are the least compact. *Trends Genet*. 2006; 22:528–32. [PubMed: 16934358]
84. Zhang MQ. Statistical features of human exons and their flanking regions. *Hum Mol Genet*. 1998; 7:919–32. [PubMed: 9536098]
85. Farlow A, Dolezal M, Hua L, Schlotterer C. The genomic signature of splicing-coupled selection differs between long and short introns. *Mol Biol Evol*. 2012; 29:21–4. [PubMed: 21878685]
86. Blanchette M, Labourier E, Green RE, Brenner SE, et al. Genome-wide analysis reveals an unexpected function for the *Drosophila* splicing factor U2AF(50) in the nuclear export of intronless mRNAs. *Mol Cell*. 2004; 14:775–86. [PubMed: 15200955]
87. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. *RNA*. 2004; 10:757–65. [PubMed: 15100430]
88. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, et al. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J*. 2004; 39:877–85. [PubMed: 15341630]
89. Pan Q, Shai O, Lee LJ, Frey BJ, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 2008; 40:1413–5. [PubMed: 18978789]
90. Zhang G, Guo G, Hu X, Zhang Y, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*. 2010; 20:646–54. [PubMed: 20305017]
91. Li P, Ponnala L, Gandotra N, Wang L, et al. The developmental dynamics of the maize leaf transcriptome. *Nat Genet*. 2010; 42:1060–7. [PubMed: 21037569]

92. Daines B, Wang H, Wang L, Li Y, et al. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. *Genome Res.* 2011; 21:315–24. [PubMed: 21177959]
93. Ramani AK, Calarco JA, Pan Q, Mavandadi S, et al. Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. *Genome Res.* 2011; 21:342–8. [PubMed: 21177968]
94. Resch AM, Palakodeti D, Lu YC, Horowitz M, et al. Transcriptome analysis reveals strain-specific and conserved stemness genes in *Schmidtea mediterranea*. *PLoS ONE.* 2012; 7:e34447. [PubMed: 22496805]
95. Tisserant E, Da Silva C, Kohler A, Morin E, et al. Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. *New Phytol.* 2011; 189:883–91. [PubMed: 21223284]
96. Wang B, Guo G, Wang C, Lin Y, et al. Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucleic Acids Res.* 2010; 38:5075–87. [PubMed: 20392818]
97. Yin Y, Yu G, Chen Y, Jiang S, et al. Genome-wide transcriptome and proteome analysis on different developmental stages of *Cordyceps militaris*. *PLoS ONE.* 2012; 7:e51853. [PubMed: 23251642]
98. Sorber K, Dimon MT, DeRisi JL. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res.* 2011; 39:3820–35. [PubMed: 21245033]
99. Xiong J, Lu X, Zhou Z, Chang Y, et al. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing. *PLoS ONE.* 2012; 7:e30630. [PubMed: 22347391]
100. Hassan MA, Melo MB, Haas B, Jensen KD, et al. De novo reconstruction of the *Toxoplasma gondii* transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. *BMC Genomics.* 2012; 13:696. [PubMed: 23231500]

Figure 1.
What happens when splicing and cleavage/polyadenylation factors compete for common or neighboring binding sites (BS) within introns and at pre-mRNA 3'-ends?

Figure 2.

An inventory of features thought to increase the local molar ratio of splicing factors (SFs) to cleavage/polyadenylation factors (CPFs) by enhancing the recruitment of splicing factors. We propose that SFs and CPFs compete for access to pre-mRNA sequence elements that reside within introns and 3' UTRs. In the proposed scenario, features of the mRNA sequence, mRNA-binding trans-acting factors (e.g. serine/arginine-rich proteins), and the ratio of phosphorylated residues in the carboxy terminal domain of the elongating RNA polymerase II serve, to different extents, to alter the spatial arrangement and the stoichiometry of CPFs and SFs. As a result, the local molar ratio of these factors in proximity to intronic binding sites (IBSs) or 3'-end termination signals is likely to change and determines whether or not events such as splicing or cleavage/polyadenylation take place.

Figure 3.

An optimally base-paired U1 snRNP exerts a distance-dependent inhibitory effect on the physical binding and/or downstream functions of cleavage/polyadenylation factors (CPFs). This effect is proposed to be widespread within introns and to influence the molar ratio between SFs and CPFs that target intronic binding sites (IBSs). **A:** In large introns, where the distance between CPF binding sites and the donor site is postulated to be large, the effect of the U1 snRNP is weak. In such cases, the access of CPFs to IBSs would be facilitated. The engagement of CPFs with IBSs would result in a less efficient recruitment of CPFs by the array of termination signals at the tail of the transcript. This would disfavor transcription yield and may facilitate alternative splicing or premature polyadenylation events if the intron in question contains key termination signals. **B:** All else being equal, in short introns, where the distance between CPF binding sites and the donor site is small, the inhibitory effect of U1 snRNP is expected to be strong. In such cases, SFs would reduce the access of CPFs to IBSs and the recruitment of CPFs by the termination signals at the pre-mRNA 3' tail would be enhanced.

Figure 4.

Average fractions (expressed in %) of alternatively spliced genes in four groups of eukaryotic species. Estimates are drawn from deep-sequencing transcriptome studies and, unless indicated otherwise, are scaled to the number of intron-containing genes in a genome. Vertebrates: *Homo sapiens* (95%; [21, 89]); land plants: *Arabidopsis thaliana* (61.2%; [47]), *Oryza sativa* (42.4%; [90]), *Zea mays* (56.4%, based on 20,999 intron-containing genes expressed in leaf tissue; [91]); invertebrates: *Drosophila melanogaster* (31%, scaled to the total number of genes in the genome; [92]), *Caenorhabditis elegans* (25%, scaled to the total number of genes in the genome; [93]), *Schmidtea mediterranea* (3.2%, scaled to the total number of genes in the genome; [94]); oligo/unicellular species: *Tuber melanosporum* (15.4%, scaled to the total number of genes in the genome; [95]), *Aspergillus oryzae* (11.1%; [96]), *Cordyceps militaris* (7.6%, scaled to the total number of genes in the genome; [97]), *Plasmodium falciparum* (8.6%, [98]), *Tetrahymena thermophila* (7.3%, the original 5.2% is rescaled to the number of intron-containing genes in the genome calculated from “final_gene_attributes.txt” retrieved from <http://ciliate.org/index.php/home/downloads>; [99]), *Toxoplasma gondii* (0.9%, [100]).

Figure 5.

Under the U1-dependent definition, the U1 snRNP binding to 5' splice sites guides splicing and influences splicing mode. **A:** Constitutive splicing of both introns a and b is promoted. Splicing of the short intron a is favored by optimal splice sites and by the (distance-dependent) inhibitory effect of the U1 snRNP on the physical binding and/or downstream functions of cleavage/polyadenylation factors (CPFs). If intron size is large (intron b), the inhibitory effect of the U1 snRNP on CPFs targeting the downstream intron binding sites (IBSs) may be reduced (indicated by red dashed cross) but splicing is still favored by the effect of the U1snRNP binding the 5' splice site of the intron downstream from exon c. In B and C, splicing of intron b is disfavored due to: **B:** the large size of intron b and/or the suboptimal binding of U1snRNP to the downstream 5' splice site; **C:** the simultaneous large size of intron b and the large size of the downstream exon c. In either case, the likelihood of skipping exon c increases. **D:** The splicing of intron d is promoted due to its short size, despite the large size of exon e.