






Methods

CATSNAPE: a user-friendly algorithm for determining the conservation of protein variants reveals extensive parallelisms in the evolution of alternative splicing

Ksenia Timofeyenko^{1,2} , Dzimtry Kanavalau³ , Panagiotis Alexiou⁴ , Maria Kalyna⁵  and Kamil Růžicka¹ 

¹Laboratory of Hormonal Regulations in Plants, Institute of Experimental Botany, Czech Academy of Sciences, 165 02 Prague 6, Czech Republic; ²Functional Genomics and Proteomics of Plants and National Centre for Biomolecular Research, Masaryk University, 625 00 Brno, Czech Republic; ³Na Vršku 15, 150 00 Prague 5, Czech Republic; ⁴Central European Institute of Technology, Masaryk University, 625 00 Brno, Czech Republic; ⁵Department of Applied Genetics and Cell Biology, Institute of Molecular Plant Biology, University of Natural Resources and Life Sciences (BOKU), 1190 Vienna, Austria

Summary

Authors for correspondence:
Ksenia Timofeyenko
Email: timofeyenko@ueb.cas.cz

Kamil Růžicka
Email: kamil.ruzicka@ueb.cas.cz

Received: 9 December 2022
Accepted: 27 January 2023

New Phytologist (2023) **238**: 1722–1732
doi: 10.1111/nph.18799

Key words: alternative splicing, bioinformatics, determinism, isoforms, machine learning, molecular evolution, transcriptome.

- Understanding the evolutionary conservation of complex eukaryotic transcriptomes significantly illuminates the physiological relevance of alternative splicing (AS). Examining the evolutionary depth of a given AS event with ordinary homology searches is generally challenging and time-consuming.
- Here, we present CATSNAP, an algorithmic pipeline for assessing the conservation of putative protein isoforms generated by AS. It employs a machine learning approach following a database search with the provided pair of protein sequences.
- We used the CATSNAP algorithm for analyzing the conservation of emerging experimentally characterized alternative proteins from plants and animals. Indeed, most of them are conserved among other species. CATSNAP can detect the conserved functional protein isoforms regardless of the AS type by which they are generated. Notably, we found that while the primary amino acid sequence is maintained, the type of AS determining the inclusion or exclusion of protein regions varies throughout plant phylogenetic lineages in these proteins. We also document that this phenomenon is less seen among animals.
- In sum, our algorithm highlights the presence of unexpectedly frequent hotspots where protein isoforms recurrently arise to carry physiologically relevant functions. The user web interface is available at <https://catsnap.cesnet.cz/>.

Introduction

In plants, animals, and other eukaryotes, alternative splicing (AS) enables the generation of multiple different mRNAs from a single gene. It is typically a major transcript that codes for a reference (canonical) protein isoform and at least one, generally less abundant, alternative splice variant. Previous studies from diverse organisms have demonstrated that AS can change properties of the resulting proteins (Stamm *et al.*, 2005; Kelemen *et al.*, 2013; Staiger & Brown, 2013; Szakonyi & Duque, 2018; Chaudhary *et al.*, 2019; Kashkan *et al.*, 2022b). A significant part of the alternative transcripts are not translated and/or are functionally neutral (Pan *et al.*, 2006; Zhang *et al.*, 2015; Tress *et al.*, 2017). However, many of them carry out relevant regulatory roles, such as control of the protein abundance via coupling with nonsense-

mediated decay (NMD) or by the timing of protein production through nuclear retention of not fully processed transcripts (Lewis *et al.*, 2003; Marquez *et al.*, 2012; Wegener & Müller-McNicoll, 2018). Hence, the biological purpose of the most AS events is obscure.

Among the main indicators of the presumed biological relevance is the evolutionary conservation of the AS event (Wang & Brendel, 2006; Keren *et al.*, 2010; Tress *et al.*, 2017). Combined transcriptomics and computational approaches have been employed to assess the conservation of AS in animals (Modrek & Lee, 2003; Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012; Xiong *et al.*, 2018) and in plants (Wang & Brendel, 2006; Severing *et al.*, 2009; Chamala *et al.*, 2015; Ling *et al.*, 2019). However, the published data frequently show several methodological limitations. For example, almost all of the computer pipelines

were designed for the identification of the conserved nucleotides flanking the area modified by the AS event and on the premise that the conserved splice isoforms are encoded by the same AS event type during evolution (Wang & Brendel, 2006; Baek *et al.*, 2008; Wang *et al.*, 2008; Darracq & Adams, 2013; Xu *et al.*, 2014; Chamala *et al.*, 2015; Mei *et al.*, 2017; Ling *et al.*, 2019). Furthermore, these reports have typically centered on a small number of representative organisms (up to 10), omitting the growing complexity of information currently available in public databases (Barbosa-Morais *et al.*, 2012; Chamala *et al.*, 2015; Mei *et al.*, 2017). In addition, a user-friendly interface determining the conservation of the provided splice isoforms is lacking. Assessing the conservation of the AS event of interest by a plain BLAST search is relatively tricky, owing to challenges in interpreting the data output. Therefore, a simple tool for performing such a task is among the experimental community highly desired.

Machine learning (ML) algorithms gained use as a powerful instrument for solving many biological questions where the given patterns can be learned on a training data set and applied to studied data, including identification of DNA and RNA protein binding motifs and prediction of splice sites (Zitnik *et al.*, 2019). Logistic regression is among the most efficient classification algorithms in ML, which achieves remarkable performance in binary classification (Lever *et al.*, 2016; Subasi, 2020).

Here we present the CATSNAP pipeline (Conserved ALternative SpliciNg in Animals and Plants). It employs a logistic regression ML model to assess the conservation of protein isoforms, comparing them to those deposited in RefSeq and GenBank. The algorithm does not take into account the type of AS event. Therefore, it is also well suited for detecting the instances where the AS events evolved several times independently with a repetitive tendency to impact equivalent protein features. A web interface dedicated to a common user is available at <https://catsnap.cesnet.cz/>.

Materials and Methods

Sequence database

The internal CATSNAP database was made by gathering protein sequences originating from genes with at least one AS event, alternative transcription start site (AltTSS) or alternative cleavage and polyadenylation in their coding regions. They were obtained from the curated RefSeq database (release 204, January 4, 2021), and extended with the complementary GenBank data set for plants (release 242.0, February 16, 2021). The full-size CATSNAP database contains sequences from 176 plants and 701 animals, and the reduced database includes 176 plant and 97 animal species (Supporting Information Table S1).

Processing of query sequences without a database accession number

For the cases when a protein sequence of interest is not present in RefSeq (animals) or RefSeq and Genbank (plants) databases, we

introduced the possibility of testing the sequences provided by the user. The algorithm requires *nucleotide* coding sequences of both isoforms (lacking the untranslated regions) and the sequence of the gene of their origin (unspliced transcript). For the user's comfort, CATSNAP removes any character different from A, T, G, C. Then, the three entered sequences are aligned by MUSCLE (Edgar, 2004) to determine the exon–intron structure, and the coding regions of the alternative isoforms are translated to amino acids. The genetic code is translated using the Standard codon table from BIOPYTHON Project (Cock *et al.*, 2009).

ML model, training set, and features

Initially, 31 conserved protein pairs from *Arabidopsis thaliana* (L.) Heynh. were selected from the literature (Table S2). They were BLASTed against the RefSeq database of plant proteins. The obtained sequences were ordered in pairs most similar to the query canonical and alternative isoforms, respectively, using the MEGA multiple sequence alignment software (Kumar *et al.*, 2018). Then, the 31 arrays containing a total of 1426 isoform pairs were used as an ML training set. To evaluate the model performance and prevent overfitting, we performed cross-validation by sequentially excluding each of the 31 initial protein pairs and all its filial hits from the training set. A logistic regression ML model from the SCIKIT library (linear_model.LogisticRegression) (Pedregosa *et al.*, 2011) was employed. Four independent ML features were specified for the ML model (Fig. S1):

- (1) The amino acid sequence similarity was set as a bit score provided by the BLAST alignment. It favors sequence pairs, which show the highest overall sequence similarity with those entered as a query.
- (2) The mutual exclusivity of the regions affected by AS (AS regions) in the sequence pair is calculated according to the formula: $F_2 = |Q \Delta H| / (|Q| + |H|)$, where Q and H are sets of positions of aligned AS regions of the query and hit sequences, respectively, and $|Q \Delta H|$ is the symmetric difference between query and hit in the AS region (Fig. S1a,b).
- (3) Amino acid similarity of AS regions is the number of matching amino acids in the AS regions in the query and hit within the subregion determined by BLAST as matching (m), divided by the length of this whole subregion (l) ($F_3 = m/l$). This feature particularly weights short conserved sequences. Identical subregions will get a score equal to 1 (Fig. S1c).
- (4) Amino acid dissimilarity shows the proportion of the matching amino acids (m) identified by the feature (3) in the context of the length of both AS regions. The formula for this feature is $F_4 = (q + h - 2m) / (q + h)$, where q is the length of the AS region of the query, h is the length of the AS region in the hit sequence, and m is the number of matches in the matching subregion. Identical subregions will get a score equal to 0 (Fig. S1d).

On the basis of the listed features, the sequence pairs receive a similarity score, which reflects the closeness of the hit pair to the query pair. The similarity score ranges from 0 to 1, where 1 is complete identity. The score is used to sort the output list of identified hits from the most similar to the least similar. For the isoforms having more than one AS region, each AS region obtains

a location identifier describing its position relative to the regions processed by constitutive splicing by counting the number of the uninterrupted constitutively spliced regions from the N- and C-terminal direction (exemplified on the Fig. S1e).

The code of the algorithm is available at GitHub (<https://github.com/kdcd/catsnap>).

Results

CATSNAPE – an ML computational pipeline for the identification of conserved AS

For assessing the conservation of a pair of isoforms of interest, the CATSNAP pipeline analyzes two protein queries. Isoform 1 is typically the reference (canonical), and Isoform 2 alternative isoform (Fig. 1a,b). They can be provided as RefSeq accession numbers or as nucleotide sequences. In the first step, both sequences are separately BLASTED against the internal database of protein variants (see the **Materials and Methods** section; Fig. 1b). As this is the most computationally demanding step of the pipeline, the user can select whether the full-sized or reduced (Table S1) database will be searched. Both BLAST output lists usually contain, besides the companion sequence and homologous isoforms, also those resulting from unrelated AS events (Fig. 1b, *DX1.3*) or AS events of paralogous genes (Fig. 1b, *CX2.1* and *CX2.2*). Next, using the RefSeq gene annotation, the algorithm separates the sequences assigned to each species and gene (Fig. 1c) and creates all possible pairwise combinations within these subsets (Fig. 1d).

To determine which of the rearranged sequence pairs are similar to the query (i.e. a conserved protein pair), the logistic regression ML model is employed. It uses pairwise alignments and scores provided by BLAST (Fig. 1e,f). The following ML features have been implemented: whole sequence amino acid similarity as determined by BLAST, (2) the position of the (non-)aligned amino acids within the AS region; and features (3) and (4) which score amino acid similarity specifically within the AS region (Fig. S1; see the **Materials and Methods** section). The candidate orthologous isoforms are returned as a file in the FASTA format, sorted from the highest to the lowest score. A large number of sequences found can complicate a quick visual examination of the results. Therefore, a list containing the single most similar isoform pair per species is also available for download and can be directly analyzed online by a built-in MUSCLE alignment tool (Fig. S2).

Plants tend to show a high degree of plasticity of the AS types during evolution

To validate the outlined algorithmic pipeline, we examined the depth of conservation of prominent experimentally validated protein isoforms from plants (Staiger & Brown, 2013; Brown *et al.*, 2015; Hrtyan *et al.*, 2015; Shang *et al.*, 2017; Szakonyi & Duque, 2018; Kashkan *et al.*, 2022b; Figs 2, S3a; Table S3). Those identified in *Arabidopsis* generally showed evolutionary conservation within Brassicales or deeper, as evidenced by a number of hits from the RefSeq and GenBank databases. This underlines that, indeed, the majority of functionally relevant protein isoforms tend

to be sustained during evolution and that CATSNAP provides a reliable baseline for assessing their conservation.

The CATSNAP pipeline is designed to assess the conservation of AS with the outcome at the protein level (Fig. 1). For example, two *Arabidopsis* isoforms of TRANSTHYRETIN-LIKE PROTEIN (TTL) differ in the presence of a peroxisome targeting signal, whose inclusion is regulated by the alternative acceptor site (AltA) in the third intron (Lamberto *et al.*, 2010). CATSNAP finds the predicted peroxisome- and cytosol-targeted isoforms in 35 eudicot and monocot species. However, in *Glycine max*, *Solanum lycopersicum*, AltA removes sole glutamate and does not affect the predicted peroxisome targeting signal, evidencing that this event is likely non-homologous in these species (Fig. S4). Hence, the evolutionary history of this event documented by Castnap illustrates the relevance of determining the conservation of AS events at the amino acid level.

From the principle, CATSNAP is able to identify the instances where the AS type varies, but the homology of the resulting amino acid sequence persists. We searched in the literature whether there were such examples presented earlier. AS of *RUBISCO ACTIVASE (RCA)* was among the first AS events identified in plants (Werneke *et al.*, 1989; Reddy, 2007). The resulting longer RCA α and shorter RCA β isoforms show a differential ability to activate the Rubisco enzyme (Zhang & Portis, 1999; Zhang *et al.*, 2002). Both proteins have been detected in multiple species, including *Arabidopsis* or various monocots (Salvucci *et al.*, 1987). To *et al.* (1999) characterized the RCA isoforms in rice. They indeed noticed that the isoforms result from a different AS type than in other species, including *Arabidopsis*. The CATSNAP algorithm, in accord with the focused RCA event evolutionary analysis (Nagarajan & Gill, 2018), revealed that the RCA isoforms are produced even by more AS types in various plants (Fig. S5). It illustrates that the protein isoforms can be functionally conserved and show high plasticity of AS types they arise from.

JASMONATE-ZIM-DOMAIN PROTEIN 10 (JAZ10) produces, besides the canonical JAZ10.1 isoform, a frame-shifted JAZ10.4 protein, which interferes with the protein interactions required for JAZ10.1 signaling. This gene also codes for the JAZ10.3 protein (from two transcripts), which impedes the JAZ10 signaling pathway in a moderate way (Chung & Howe, 2009; Moreno *et al.*, 2013; Fig. S6a). CATSNAP detects the JAZ10.4 orthologs in multiple species within the Brassicaceae family, being the products of the same AS type. JAZ10.3 orthologs are found in eudicots and monocots. However, the CATSNAP search revealed that the types of AS in *JAZ10.3* highly vary among other plants (Figs 2a, S6b,c). Moreover, SGR5 β , a truncated non-DNA binding isoform of the transcriptional factor SHOOT GRAVITROPISM 5 (SGR5) (Kim *et al.*, 2016), can be seen in the wide range of plant species, including the liverwort *Marchantia polymorpha*. We observed, too, that the different AS types and AltTSSs lead to the production of the proteins matching the alternative SGR5 β isoform (Fig. S7). Thus, the CATSNAP pipeline efficiently allows for detecting the instances when the isoforms are functionally conserved at the amino acid level, but underlying mechanisms at the nucleotide level (AS types or AltTSS) can differ during gene evolution.

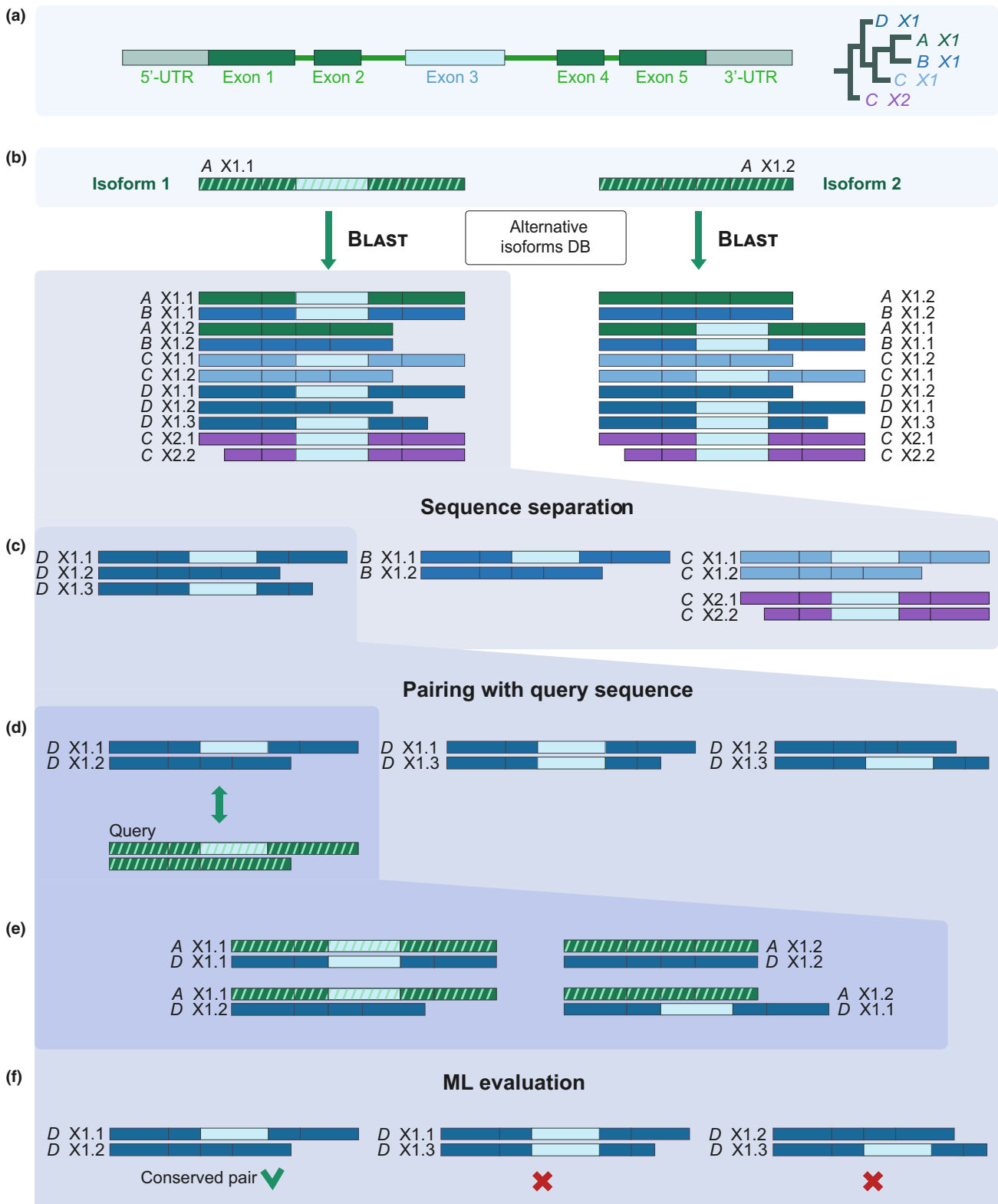


Fig. 1 The scheme of the CATSNAP algorithmic pipeline. (a) A diagram of an example gene X1 with alternative skipping of exon 3 (turquoise); the outline of its imaginary phylogenetic relationships (from Organism A to D) is on the right. (b) The query isoforms A X1.1 and A X1.2 (hatched) are BLASTed against the internal database of protein isoforms. The BLAST output also contains isoforms originating from unrelated alternative splicing (AS) events (D X1.3) or occurring in paralogous genes (C X2). (c) The protein isoforms identified by BLAST are separated according to the gene name and organismal source. (d) Protein isoforms from each organism are paired by all combinations. (e) Each generated pair is re-associated with the query sequences using initial BLAST alignments (c), and, (f) evaluated.

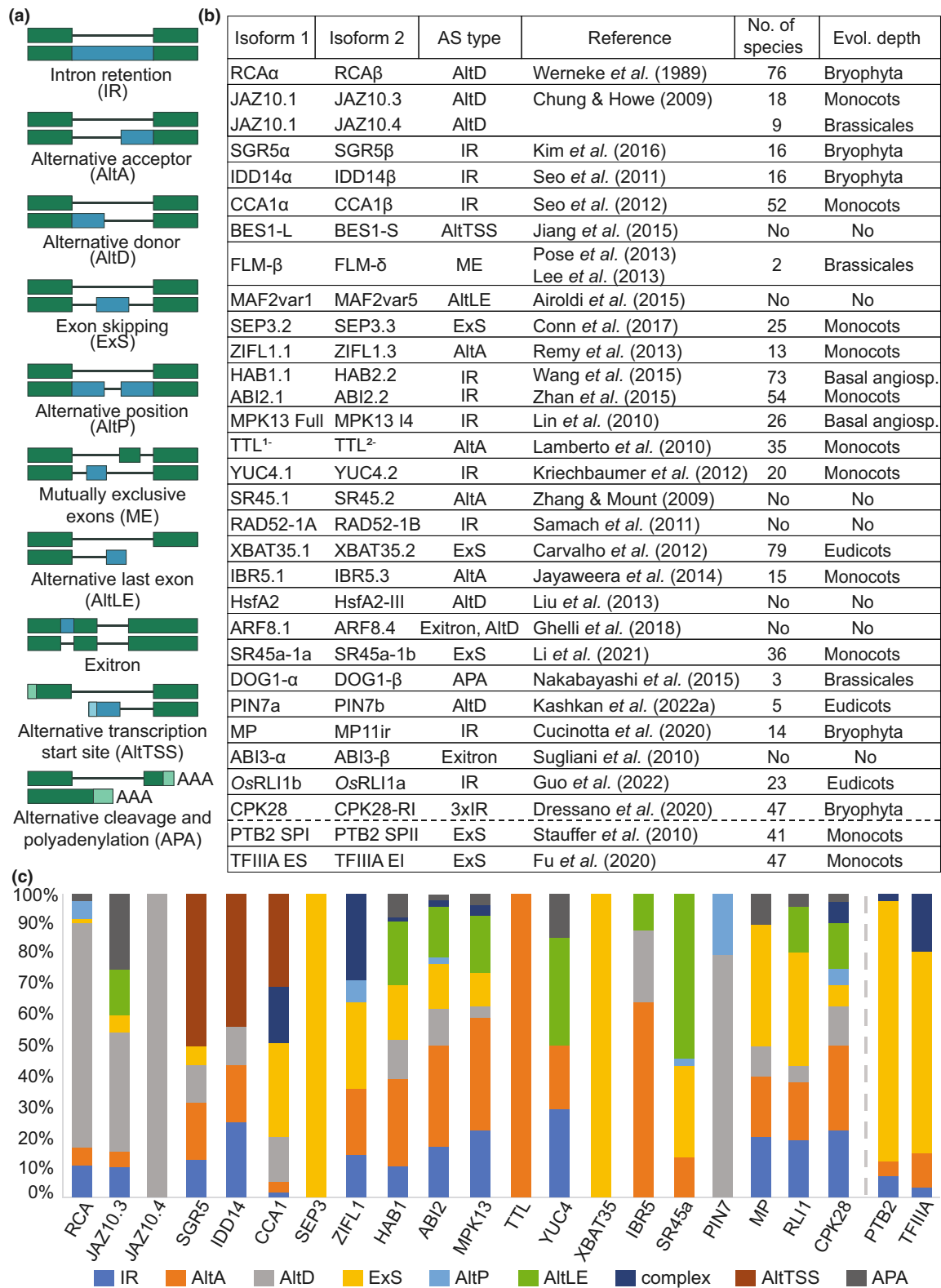


Fig. 2 Evolutionary conservation of functionally validated alternative isoforms from plants. (a) The schemes of relevant alternative splicing (AS) types, in addition to the alternative transcription start site (AltTSS) and alternative polyadenylation (APA). Dark green and dark blue colors represent coding constitutive and alternative regions, respectively; light green and light blue colors are for non-coding constitutive and alternative regions, respectively. (b) The conservation of validated *Arabidopsis thaliana* and *Oryza sativa* (Os) isoforms within main plant phylogenetic lineages, including example transcripts without protein-coding potential (undergoing nonsense-mediated decay (NMD), separated by dashed line). (c) In most of the functionally validated plant AS events, the changes seen at the amino acid level are widely conserved, independent of the AS type (or AltTSS and APA), it even includes the transcripts undergoing NMD (separated by dashed line); the y-axis denotes the proportions of alternative isoforms arising from the different outlined mRNA processing type.

CALCIUM-DEPENDENT PROTEIN KINASE 28 (CPK28) produces an alternative isoform with a premature termination codon (PTC), predicted to remove Ca^{2+} -binding EF-hands domains, required for the kinase activity of the resulting protein (Dressano *et al.*, 2020). This is likely a characteristic of the broader CPK family (CDPK) and was suggested as exerted via several AS types among angiosperms (Loranger *et al.*, 2021). Accordingly, CATSNAP recognizes the conservation of the shortened CPK28-RI isoforms up to bryophytes (Fig. S8). Interestingly, the presented molecular model (Dressano *et al.*, 2020) does not exclude, in principle, the scenario that the *CPK28-RI* transcript might actually not be translated. Hence, we also examined the transcripts, which have earlier been shown to be controlled by AS coupled with NMD. They are associated with PTCs; they are not translated and undergo subsequent degradation (Lewis *et al.*, 2003). Among them, the NMD-controlled autoregulatory circuits of the polypyrimidine tract-binding proteins (PTBs) have been intensively investigated (Wolterton *et al.*, 2004; Stauffer *et al.*, 2010). Indeed, the sequences related to the imaginary truncated protein derived from the *Arabidopsis* alternative *PTB2 SPII* transcript were found in 41 plant species up to monocots (Fig. S9). Similarly, the expression of *TRANSCRIPTION FACTOR FOR POLYMERASE III A (TFIIIA)* is autoregulated via the generation of an alternative NMD-dependent transcript (Fu *et al.*, 2009). CATSNAP finds the nominal shortened proteins in 25 angiosperms (Fig. S10). Thus, CATSNAP is able to detect the conservation of a regulatory event (even) regardless of the presumed isoform translatability.

As it appears that many plant genes show a high degree of plasticity of the AS types during evolution, we systematically inspected the types of AS in the protein sequences related to the remaining experimentally validated AS events in plants (Fig. 2b). Strikingly, we found that the variability of AS types is seen among the most well-characterized alternative isoforms (Fig. 2c). Altogether, we conclude that examining the evolutionary conservation on the basis of amino acid sequence, as provided by

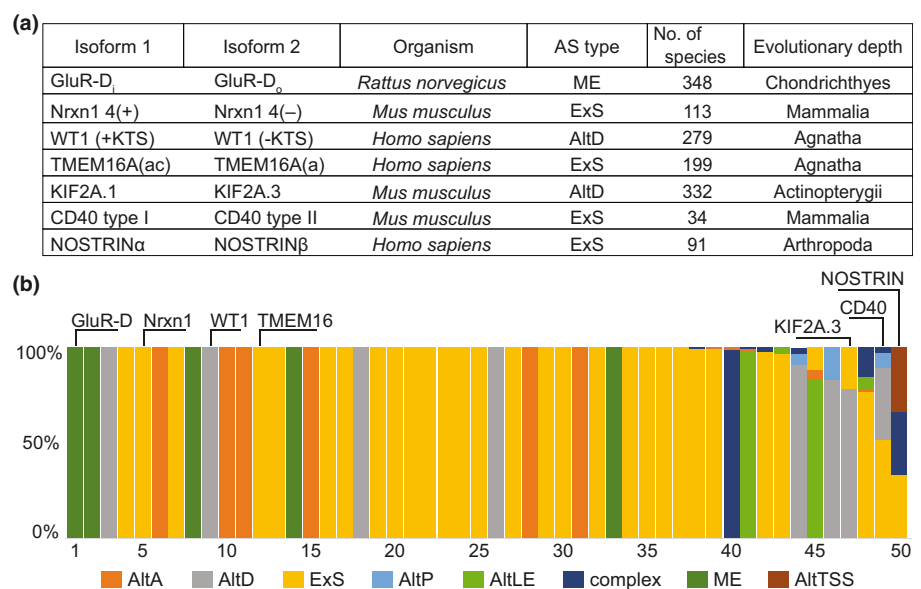
CATSNAP, brings a remarkable insight into the evolutionary conservation of AS. The homologous protein isoforms are maintained during evolution, but the underlying AS types can largely vary with respect to the organismal group.

The evolution and plasticity of AS in animals

To test the versatility of the CATSNAP pipeline further, we also examined the conservation of prominent characterized AS events in animals (Stamm *et al.*, 2005; Kelemen *et al.*, 2013). For example, for the GluR-D_i and GluR-D_o isoforms of the AMPA-type ionotropic glutamate receptors (Sommer *et al.*, 1990; Mosbacher *et al.*, 1994; Dawe *et al.*, 2019; Zhao *et al.*, 2019), we found that both are conserved among > 300 vertebrate species ranging from the cartilage fishes (Chondrichthyes) to mammals (Figs S3b, S11). Furthermore, the CATSNAP pipeline documented a stable and long evolutionary history of selected prominent animal events, including that of neurexin I (*Nrxn1*) (Iijima *et al.*, 2011), the Wilms tumor susceptibility gene (*WT1*) (Larsson *et al.*, 1995), transmembrane 16A (*TMEM16A* or *anoctamin1*) (Ko *et al.*, 2020; Fig. 3a), and a large number of other functionally confirmed isoforms (Figs 3b, S3b; Table S4). Altogether, this demonstrates that although the CATSNAP ML algorithm has been trained on plant sequences (see the Materials and Methods section), it can also be used for examining the evolutionary history of protein isoforms in other eukaryotes, including animals.

We analyzed the AS plasticity of the experimentally validated AS events from animals, too (Fig. 3b). The KIF2A.3 isoform of the microtubule destabilizer Kinesin family member 2A lacks 20 amino acids by the alternative choice of a donor splice site (AltD) in the 5th intron (Akkaya *et al.*, 2021; Figs 3, S12a). The alternative inclusion of the internal protein motif is conserved up to the ancient teleost fishes. However, in the evolutionarily derived bony fishes (Euteleostei), this region is encoded by a separate exon, and the event is regulated by exon skipping (Fig. S12b–d). The alternative isoform of tumor necrosis factor receptor CD40 lacks a C-

Fig. 3 Conservation of functionally validated alternative isoforms from various animal model organisms. (a) The summary of the evolutionary history of validated animal protein isoforms discussed in the main text. (b) Some validated alternative splicing (AS) events in animals show evolutionary plasticity; several detected rare changes of AS type can be perhaps ascribed to the misannotation of the corresponding transcript. The values on the x-axis indicate the analyzed isoform pairs listed in Supporting Information Table S4; the y-axis denotes the proportions of alternative isoforms arising from different AS types, alternative transcription start site (ALTSS) and alternative polyadenylation (APA).



terminal part present in the canonical protein, including a transmembrane domain (Tone *et al.*, 2000; Eshel *et al.*, 2008; Hou *et al.*, 2008). The truncated version is conserved in mammals and is processed by multiple AS types as well (Fig. S13). Human NOS-TRIN (eNOS trafficking inducer) undergoes shortening of its N-terminus in the stressed liver, resulting in the NOS-TRIN β isoform (Mookerjee *et al.*, 2007; Wiesenthal *et al.*, 2009). NOS-TRIN β is conserved up to cartilaginous fishes and arthropods and arises from exon skipping, various combinations of AS types and AltTSSs (Fig. S14). Taken together, although much less prevalent than in plants (compare Figs 2c, 3b), the evolutionary plasticity of AS types is also seen in animal systems.

Discussion

Several studies demonstrated that evolutionary conservation closely correlates with the functionality of a given isoform (McCartney *et al.*, 2005; Lamberto *et al.*, 2010; Astro *et al.*, 2022). On the contrary, another set of reports has presented that mutations underlying AS are linked with recent evolutionary adaptations, highlighting the transitivity of AS events (Barbosa-Morais *et al.*, 2012; Ling *et al.*, 2019; Wright *et al.*, 2022). To address the two extreme viewpoints, as a proof of concept of the algorithm, we show here that the functionally validated isoforms arising from AS tend to exhibit prevalently deeper evolutionary origin. Moreover, we have designed a user-friendly interface, available at <https://catsnap.cesnet.cz/>, which allows, even in a batch mode, for a quick visual overview of the evolutionary conservation of protein (transcript) variants of choice. Together with other available large-scale data resources (e.g. Berardini *et al.*, 2015; Martín *et al.*, 2021; Cunningham *et al.*, 2022; Gramates *et al.*, 2022), it is aimed to help the researcher to hint at whether the event of interest could have a detectable biological function, and be, for instance, suitable for further experimental characterization.

In contrast to the effort done previously (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012; Darracq & Adams, 2013; Chalmala *et al.*, 2015; Ling *et al.*, 2019), we employed an amino acid sequence view on the conservation of AS. Our approach reveals that the expression of the conserved functional isoforms is in different plant species likely commonly controlled by different types of AS. It should be underlined that the protein models present in the current databases mostly arise from algorithmic predictions. The annotated proteins may show a different authentic amino acid sequence or might not be translated at all. Thus, the results should be interpreted critically (Brown *et al.*, 2015). On the contrary, from previous reports, the two RCA isoforms, processed by multiple types of AS (Figs 2c, S5; Nagarajan & Gill, 2018), have been functionally characterized in different plant species (Werneke *et al.*, 1989; To *et al.*, 1999; Xu *et al.*, 2017). A similar phenomenon has recently been proposed for the isoforms of REGULATOR OF LEAF INCLINATION 1 (RLI1a and b) in rice and *Arabidopsis*, where they show a high evolutionary plasticity of the AS events controlling their expression (Fig. 2c; Guo *et al.*, 2022). Moreover, the individual RCA isoforms can even be encoded by separate genes in several species (Salvucci *et al.*, 2003; Yin *et al.*, 2010; Nagarajan & Gill, 2018) and the differentially

localized auxin synthase isoforms YUCCA4 (YUC4) are parallelized by individual gene products of the YUC family in *Arabidopsis* (Fig. 2c; Kriebchaumer *et al.*, 2012, 2016). Taken together, it seems that particularly plant genes show a high extent of evolutionary plasticity of protein isoforms controlled by AS.

In contrast to the situation in animals, plants show high variability in genome sizes (explicitly in terms of ploidy), general DNA organization and sequence divergence. In addition, their coding and non-coding sequences evolve faster (Leitch & Leitch, 2008; Kejnovsky *et al.*, 2009; Murat *et al.*, 2012). Our analysis of the experimentally validated alternative isoforms revealed that the AS patterns in plants broadly vary compared to animals as well. Regulation of AS is jointly carried out by the *cis*-elements, encoded by the pre-mRNA sequence, and *trans*-acting protein regulators, which bind these motifs (Reddy, 2007; Fu & Ares, 2014). Ling *et al.* (2019) noted a considerably high rate of gains and losses of AS among plant transcriptomes analyzed, strongly linked with the rapid evolution of plant *cis*-elements (Shen *et al.*, 2014; Thatcher *et al.*, 2014; Ling *et al.*, 2019; Wang *et al.*, 2019). Altogether, the presented plasticity of the prominent protein isoforms in plants represents just another manifestation of the remarkable variability of their genomes. Hence, the conservation of the protein isoforms can mark a hotspot that leads to the production of the same evolutionary conserved regulators and recurrent functional adaptation.

Acknowledgements

We thank Nicholas Provart for his comments on the user interface of the algorithm, Elena Zemlyanskaya and Ivan Kashkan for discussions. This work was supported by the Czech Science Foundation (23-08067S) to KT and KR, and the Ministry of Education, Youth and Sports of the Czech Republic – Centre for Experimental Plant Biology (CZ.02.1.01/0.0/0.0/16_019/0000738) to KR, e-INFRA CZ (ID: 90140) to CESNET, Czech National Academic e-Infrastructure Association, and Austrian Science Fund (I 3551) to MK.

Competing interests

None declared.

Author contributions

KT, DK, PA, MK, and KR conceptualized the research. KT and DK designed the algorithm. KT, MK and KR analyzed the data. KT and KR wrote the manuscript. All authors read and approved the final version of the manuscript.

ORCID

Panagiotis Alexiou  <https://orcid.org/0000-0003-3437-7482>
 Maria Kalyna  <https://orcid.org/0000-0003-4702-7625>
 Dzimtry Kanavalau  <https://orcid.org/0000-0002-4802-1631>
 Kamil Růžička  <https://orcid.org/0000-0002-0602-2046>
 Ksenia Timofeyenko  <https://orcid.org/0000-0002-1454-862X>

Data availability

The code of the algorithm is available at GitHub (<https://github.com/kdcd/catsnap>).

References

- Airoidi CA, McKay M, Davies B. 2015. *MAF2* is regulated by temperature-dependent splicing and represses flowering at low temperatures in parallel with *FLM*. *PLoS ONE* 10: e0126516.
- Akkaya C, Atak D, Kamacioglu A, Akarlar BA, Guner G, Bayam E, Taskin AC, Ozlu N, Ince-Dunn G. 2021. Roles of developmentally regulated KIF2A alternative isoforms in cortical neuron migration and differentiation. *Development* 148: dev192674.
- Astro V, Ramirez-Calderon G, Pennucci R, Caroli J, Saera-Vila A, Cardona-Londoño K, Forastieri C, Fiacco E, Maksoud F, Aloysi M *et al.* 2022. Fine-tuned KDM1A alternative splicing regulates human cardiomyogenesis through an enzymatic-independent mechanism. *iScience* 25: 104665.
- Baek J-M, Han P, Iandolino A, Cook DR. 2008. Characterization and comparison of intron structure and alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, *Arabidopsis* and rice. *Plant Molecular Biology* 67: 499–510.
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Guerousov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R *et al.* 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587–1593.
- Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *Genesis* 53: 474–485.
- Brown JWS, Simpson CG, Marquez Y, Gadd GM, Barta A, Kalyna M. 2015. Lost in translation: pitfalls in deciphering plant alternative splicing transcripts. *Plant Cell* 27: 2083–2087.
- Carvalho SD, Saraiva R, Maia TM, Abreu IA, Duque P. 2012. XBAT35, a novel *Arabidopsis* RING E3 ligase exhibiting dual targeting of its splice isoforms, is involved in ethylene-mediated regulation of apical hook curvature. *Molecular Plant* 5: 1295–1309.
- Chamala S, Feng G, Chavarro C, Barbazuk WB. 2015. Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Frontiers in Bioengineering and Biotechnology* 3: 33.
- Chaudhary S, Khokhar W, Jabre I, Reddy ASN, Byrne LJ, Wilson CM, Syed NH. 2019. Alternative splicing and protein diversity: plants versus animals. *Frontiers in Plant Science* 10: 708.
- Chung HS, Howe GA. 2009. A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in *Arabidopsis*. *Plant Cell* 21: 131–145.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al.* 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423.
- Conn VM, Hugouvieux V, Nayak A, Conos SA, Capovilla G, Cildir G, Jourdain A, Tergaonkar V, Schmid M, Zubieta C *et al.* 2017. A circRNA from *SEPALLATA3* regulates splicing of its cognate mRNA through R-loop formation. *Nature Plants* 3: 17053.
- Cucinotta M, Cavalleri A, Guazzotti A, Astori C, Manrique S, Bombarely A, Oliveto S, Biffo S, Weijers D, Kater MM *et al.* 2020. Alternative splicing generates a MONOPTEROS isoform required for ovule development. *Current Biology* 31: 892–899.
- Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R *et al.* 2022. Ensembl 2022. *Nucleic Acids Research* 50: D988–D995.
- Darracq A, Adams KL. 2013. Features of evolutionarily conserved alternative splicing events between *Brassica* and *Arabidopsis*. *New Phytologist* 199: 252–263.
- Dawe GB, Kadir MF, Venskutonytė R, Perozzo AM, Yan Y, Alexander RPD, Navarrete C, Santander EA, Arsenaault M, Fuentes C *et al.* 2019. Nanoscale mobility of the apo state and TARP stoichiometry dictate the gating behavior of alternatively spliced AMPA receptors. *Neuron* 102: 976–992.
- Dressano K, Weckwerth PR, Poretzky E, Takahashi Y, Villarreal C, Shen Z, Schroeder JI, Briggs SP, Huffaker A. 2020. Dynamic regulation of Pep-induced immunity through post-translational control of defence transcript splicing. *Nature Plants* 6: 1008–1019.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.
- Eshel D, Toporik A, Efrati T, Nakav S, Chen A, Douvdevani A. 2008. Characterization of natural human antagonistic soluble CD40 isoforms produced through alternative splicing. *Molecular Immunology* 46: 250–257.
- Fu X-D, Ares M. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* 15: 689–701.
- Fu Y, Bannach O, Chen H, Teune J-H, Schmitz A, Steger G, Xiong L, Barbazuk WB. 2009. Alternative splicing of anciently exozoned 5S rRNA regulates plant transcription factor TFIIIA. *Genome Research* 19: 913–921.
- Ghelli R, Brunetti P, Napoli N, De Paolis A, Cecchetti V, Tsuge T, Serino G, Matsui M, Mele G, Rinaldi G *et al.* 2018. A newly identified flower-specific splice variant of *AUXIN RESPONSE FACTOR8* regulates stamen elongation and endothecium lignification in *Arabidopsis*. *Plant Cell* 30: 620–637.
- Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T *et al.* 2022. FlyBase: a guided tour of highlighted features. *Genetics* 220: iyac035.
- Guo M, Zhang Y, Jia X, Wang X, Zhang Y, Liu J, Yang Q, Ruan W, Yi K. 2022. Alternative splicing of *REGULATOR OF LEAF INCLINATION 1* modulates phosphate starvation signaling and growth in plants. *Plant Cell* 34: 3319–3338.
- Hou H, Obregon D, Lou D, Ehrhart J, Fernandez F, Silver A, Tan J. 2008. Modulation of neuronal differentiation by CD40 isoforms. *Biochemical and Biophysical Research Communications* 369: 641–647.
- Hrtyan M, Šliková E, Hejčík J, Růžická K. 2015. RNA processing in auxin and cytokinin pathways. *Journal of Experimental Botany* 66: 4897–4912.
- Iijima T, Wu K, Witte H, Hanno-Iijima Y, Glatter T, Richard S, Scheiffele P. 2011. SAM68 regulates neuronal activity-dependent alternative splicing of neurexin-1. *Cell* 147: 1601–1614.
- Jayaweera T, Siriwardana C, Dharmasiri S, Quint M, Gray WM, Dharmasiri N. 2014. Alternative splicing of *Arabidopsis* IBR5 pre-mRNA generates two IBR5 isoforms with distinct and overlapping functions. *PLoS ONE* 9: e102301.
- Jiang J, Zhang C, Wang X. 2015. A recently evolved isoform of the transcription factor BES1 promotes brassinosteroid signaling and development in *Arabidopsis thaliana*. *Plant Cell* 27: 361–374.
- Kashkan I, Hrtyan M, Retzer K, Humpolíčková J, Jayasree A, Filepová R, Vondráková Z, Simon S, Rombaut D, Jacobs TB *et al.* 2022a. Mutually opposing activity of PIN7 splicing isoforms is required for auxin-mediated tropic responses in *Arabidopsis thaliana*. *New Phytologist* 233: 329–343.
- Kashkan I, Timofeyenko K, Růžická K. 2022b. How alternative splicing changes the properties of plant proteins. *Quantitative Plant Biology* 3: 1–11.
- Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution* 24: 572–582.
- Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. *Gene* 514: 1–30.
- Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* 11: 345–355.
- Kim J-Y, Ryu JY, Baek K, Park C-M. 2016. High temperature attenuates the gravitropism of inflorescence stems by inducing *SHOOT GRAVITROPISM 5* alternative splicing in *Arabidopsis*. *New Phytologist* 209: 265–279.
- Ko W, Jung S-R, Kim K-W, Yeon J-H, Park C-G, Nam JH, Hille B, Suh B-C. 2020. Allosteric modulation of alternatively spliced Ca²⁺-activated Cl⁻ channels TMEM16A by PI(4,5)P2 and CaMKII. *Proceedings of the National Academy of Sciences, USA* 117: 30787–30798.
- Kriebchaumer V, Botchway SW, Hawes C. 2016. Localization and interactions between *Arabidopsis* auxin biosynthetic enzymes in the TAA/YUC-dependent pathway. *Journal of Experimental Botany* 67: 4195–4207.
- Kriebchaumer V, Wang P, Hawes C, Abell BM. 2012. Alternative splicing of the auxin biosynthesis gene *YUCCA4* determines its subcellular compartmentation. *The Plant Journal* 70: 292–302.

- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. Mega X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* 35: 1547–1549.
- Lamberto I, Percudani R, Gatti R, Folli C, Petrucco S. 2010. Conserved alternative splicing of *Arabidopsis* transthyretin-like determines protein localization and S-allantoin synthesis in peroxisomes. *Plant Cell* 22: 1564–1574.
- Larsson SH, Miyagawa K, Engelkamp D, Rassoulzadegan M, Ross A, Cuzin F, Hastie ND. 1995. Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. *Cell* 81: 391–401.
- Lee JH, Ryu H-S, Chung KS, Pose D, Kim S, Schmid M, Ahn JH. 2013. Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. *Science* 342: 628–632.
- Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* 320: 481–483.
- Lever J, Krzywinski M, Altman N. 2016. Logistic regression. *Nature Methods* 13: 541–542.
- Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences, USA* 100: 189–192.
- Li Y, Guo Q, Liu P, Huang J, Zhang S, Yang G, Wu C, Zheng C, Yan K. 2021. Dual roles of the serine/arginine-rich splicing factor SR45a in promoting and interacting with nuclear cap-binding complex to modulate the salt-stress response in *Arabidopsis*. *New Phytologist* 230: 641–655.
- Lin W-Y, Matsuoka D, Sasayama D, Nanmori T. 2010. A splice variant of *Arabidopsis* mitogen-activated protein kinase and its regulatory function in the MKK6–MPK13 pathway. *Plant Science* 178: 245–250.
- Ling Z, Brockmüller T, Baldwin IT, Xu S. 2019. Evolution of alternative splicing in eudicots. *Frontiers in Plant Science* 10: 707.
- Liu J, Sun N, Liu M, Liu J, Du B, Wang X, Qi X. 2013. An autoregulatory loop controlling *Arabidopsis HsfA2* expression: role of heat shock-induced alternative splicing. *Plant Physiology* 162: 512–521.
- Loranger MEW, Huffaker A, Monaghan J. 2021. Truncated variants of Ca^{2+} -dependent protein kinases: a conserved regulatory mechanism? *Trends in Plant Science* 26: 1002–1005.
- Marquez Y, Brown JWS, Simpson C, Barta A, Kalyna M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Research* 22: 1184–1195.
- Martín G, Márquez Y, Mantica F, Duque P, Irimia M. 2021. Alternative splicing landscapes in *Arabidopsis thaliana* across tissues and stress conditions highlight major functional differences with animals. *Genome Biology* 22: 35.
- McCartney CE, McClafferty H, Huibant J-M, Rowan EG, Shipston MJ, Rowe ICM. 2005. A cysteine-rich motif confers hypoxia sensitivity to mammalian large conductance voltage- and Ca-activated K (BK) channel α -subunits. *Proceedings of the National Academy of Sciences, USA* 102: 17870–17876.
- Mei W, Boatwright L, Feng G, Schnable JC, Barbazuk WB. 2017. Evolutionarily conserved alternative splicing across monocots. *Genetics* 207: 465–480.
- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338: 1593–1599.
- Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics* 34: 177–180.
- Mookerjee RP, Wiesenthal A, Icking A, Hodges SJ, Davies NA, Schilling K, Sen S, Williams R, Novelli M, Müller-Esterl W *et al.* 2007. Increased gene and protein expression of the novel eNOS regulatory protein NOSTRIN and a variant in alcoholic hepatitis. *Gastroenterology* 132: 2533–2541.
- Moreno JE, Shyu C, Campos ML, Patel LC, Chung HS, Yao J, He SY, Howe GA. 2013. Negative feedback control of jasmonate signaling by an alternative splice variant of JAZ101. *Plant Physiology* 162: 1006–1017.
- Mosbacher J, Schoepfer R, Monyer H, Burnashev N, Seeburg PH, Ruppersberg JP. 1994. A molecular determinant for submillisecond desensitization in glutamate receptors. *Science* 266: 1059–1062.
- Murat F, de Peer YV, Salse J. 2012. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biology and Evolution* 4: 917–928.
- Nagarajan R, Gill KS. 2018. Evolution of Rubisco activase gene in plants. *Plant Molecular Biology* 96: 69–87.
- Nakabayashi K, Bartsch M, Ding J, Soppe WJJ. 2015. Seed dormancy in *Arabidopsis* requires self-binding ability of DOG1 protein and the presence of multiple isoforms generated by alternative splicing. *PLoS Genetics* 11: 1–20.
- Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes & Development* 20: 153–158.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al.* 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.
- Posé D, Verhage L, Ott F, Yant L, Mathieu J, Angenent GC, Immink RGH, Schmid M. 2013. Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* 503: 414–417.
- Reddy ASN. 2007. Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annual Review of Plant Biology* 58: 267–294.
- Remy E, Cabrito TR, Baster P, Batista RA, Teixeira MC, Friml J, Sá-Correia I, Duque P. 2013. A major facilitator superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in *Arabidopsis*. *Plant Cell* 25: 901–926.
- Salvucci ME, van de Loo FJ, Stecher D. 2003. Two isoforms of Rubisco activase in cotton, the products of separate genes not alternative splicing. *Planta* 216: 736–744.
- Salvucci ME, Werneke JM, Ogren WL, Portis AR. 1987. Purification and species distribution of Rubisco activase. *Plant Physiology* 84: 930–936.
- Samach A, Melamed-Bessudo C, Avivi-Ragolski N, Pietrokovski S, Levy AA. 2011. Identification of plant *RAD52* homologs and characterization of the *Arabidopsis thaliana RAD52*-like genes. *Plant Cell* 23: 4266–4279.
- Seo PJ, Kim MJ, Ryu J-Y, Jeong E-Y, Park C-M. 2011. Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. *Nature Communications* 2: 303.
- Seo PJ, Park M-J, Lim M-H, Kim S-G, Lee M, Baldwin IT, Park C-M. 2012. A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in *Arabidopsis*. *Plant Cell* 24: 2427–2442.
- Severing EI, van Dijk AD, Stiekema WJ, van Ham RC. 2009. Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics* 10: 154.
- Shang X, Cao Y, Ma L. 2017. Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. *International Journal of Molecular Sciences* 18: 432.
- Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong L-A, Peng D-L *et al.* 2014. Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 26: 996–1008.
- Sommer B, Keinänen K, Verdoorn TA, Wisden W, Burnashev N, Herb A, Kohler M, Takagi T, Sakmann B, Seeburg PH. 1990. Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* 249: 1580–1585.
- Staiger D, Brown JWS. 2013. Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* 25: 3640–3656.
- Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* 344: 1–20.
- Stauffer E, Westermann A, Wagner G, Wachter A. 2010. Polypyrimidine tract-binding protein homologues from *Arabidopsis* underlie regulatory circuits based on alternative splicing and downstream control. *The Plant Journal* 64: 243–255.
- Subasi A. 2020. Chapter 3 – Machine learning techniques. In: Trombaco RG, ed. *Practical machine learning for data analysis using Python*. Cambridge, MA, USA: Academic Press, 91–202.
- Sugliani M, Brambilla V, Clercx EJM, Koornneef M, Soppe WJJ. 2010. The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in *Arabidopsis*. *Plant Cell* 22: 1936–1946.

- Szakonyi D, Duque P. 2018. Alternative splicing as a regulator of early plant development. *Frontiers in Plant Science* 9: 1174.
- Thatcher SR, Zhou W, Leonard A, Wang B-B, Beatty M, Zastrow-Hayes G, Zhao X, Baumgarten A, Li B. 2014. Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. *Plant Cell* 26: 3472–3487.
- To K-Y, Suen D-F, Chen S-CG. 1999. Molecular characterization of ribulose-1,5-bisphosphate carboxylase/oxygenase activase in rice leaves. *Planta* 209: 66–76.
- Tone M, Tone Y, Fairchild P, Wykes M, Waldmann H. 2000. Regulation of CD40 function by its isoforms generated through alternative splicing. *Proceedings of the National Academy of Sciences, USA* 98: 1751–1756.
- Tress ML, Abascal F, Valencia A. 2017. Alternative splicing may not be the key to proteome complexity. *Trends in Biochemical Sciences* 42: 98–110.
- Wang B-B, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences, USA* 103: 7175–7180.
- Wang B-B, O'Toole M, Brendel V, Young ND. 2008. Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biology* 8: 17.
- Wang X, Yang M, Ren D, Terzaghi W, Deng X-W, He G. 2019. Cis-regulated alternative splicing divergence and its potential contribution to environmental responses in Arabidopsis. *The Plant Journal* 97: 555–570.
- Wang Z, Ji H, Yuan B, Wang S, Su C, Yao B, Zhao H, Li X. 2015. ABA signalling is fine-tuned by antagonistic HAB1 variants. *Nature Communications* 6: 8138.
- Wegener M, Müller-McNicoll M. 2018. Nuclear retention of mRNAs – quality control, gene regulation and human disease. *Seminars in Cell & Developmental Biology* 79: 131–142.
- Werneke JM, Chatfield JM, Ogren WL. 1989. Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis. *Plant Cell* 1: 815–825.
- Wiesenthal A, Hoffmeister M, Siddique M, Kovacevic I, Oess S, Müller-Esterl W, Siehoff-Icking A. 2009. NOSTRIN β – a shortened NOSTRIN variant with a role in transcriptional regulation. *Traffic* 10: 26–34.
- Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CWJ. 2004. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Molecular Cell* 13: 91–100.
- Wright CJ, Smith CWJ, Jiggins CD. 2022. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics* 23: 697–710.
- Xiong J, Jiang X, Ditsiou A, Gao Y, Sun J, Lowenstein ED, Huang S, Khaitovich P. 2018. Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages. *Human Molecular Genetics* 27: 1474–1485.
- Xu P, Kong Y, Song D, Huang C, Li X, Li L. 2014. Conservation and functional influence of alternative splicing in wood formation of *Populus* and *Eucalyptus*. *BMC Genomics* 15: 780.
- Xu S, Qin Z-Y, Gong P-C, Dong Q-L, Bao Y. 2017. Identification and characterization of Rubisco activase genes in *Oryza punctata*. *Journal of Systematics and Evolution* 55: 200–207.
- Yin Z, Meng F, Song H, Wang X, Xu X, Yu D. 2010. Expression quantitative trait loci analysis of two genes encoding Rubisco activase in soybean. *Plant Physiology* 152: 1625–1637.
- Zhan X, Qian B, Cao F, Wu W, Yang L, Guan Q, Gu X, Wang P, Okusolubo TA, Dunn SL *et al.* 2015. An Arabidopsis PWI and RRM motif-containing protein is critical for pre-mRNA splicing and ABA responses. *Nature Communications* 6: 8139.
- Zhang N, Kallis RP, Ewy RG, Portis AR. 2002. Light modulation of Rubisco in Arabidopsis requires a capacity for redox regulation of the larger Rubisco activase isoform. *Proceedings of the National Academy of Sciences, USA* 99: 3330–3334.
- Zhang N, Portis AR. 1999. Mechanism of light regulation of Rubisco: a specific role for the larger Rubisco activase isoform involving reductive activation by thioredoxin-f. *Proceedings of the National Academy of Sciences, USA* 96: 9438–9443.
- Zhang X, Rosen BD, Tang H, Krishnakumar V, Town CD. 2015. Polyribosomal RNA-seq reveals the decreased complexity and diversity of the Arabidopsis translatoome. *PLoS ONE* 10: e0117699.
- Zhang X-N, Mount SM. 2009. Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. *Plant Physiology* 150: 1450–1458.
- Zhao Y, Chen S, Swensen AC, Qian W-J, Gouaux E. 2019. Architecture and subunit arrangement of native AMPA receptors elucidated by cryo-EM. *Science* 364: 355–362.
- Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. 2019. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Information Fusion* 50: 71–91.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 The outline of the CATSNAP machine learning features.

Fig. S2 The snapshots of the Catnap graphical output interface.

Fig. S3 Schematic relationships of the main plant and animal phylogenetic groups.

Fig. S4 Alternative splicing of *TTL* from representative plant species.

Fig. S5 Alternative splicing of *RCA* in various plants.

Fig. S6 Alternative splicing of *JAZ10* in various plants.

Fig. S7 Alternative splicing and alternative transcription start sites of *SGR5* in various plants.

Fig. S8 Alternative splicing of *CPK28* in various plants.

Fig. S9 Alternative splicing of *PTB2* in various plants.

Fig. S10 Alternative splicing of *TFIIIA* in various plants.

Fig. S11 Alternative splicing of *Glu4* in various animals.

Fig. S12 Alternative splicing of *Kif2a* in various animals.

Fig. S13 Alternative splicing of *CD40* in various animals.

Fig. S14 Alternative splicing and alternative transcription start sites of various animal *NOSTRIN* genes.

Table S1 Animal species included in the reduced web-mode database of alternative isoforms.

Table S2 Conserved *Arabidopsis thaliana* alternative splicing events used as an initial source for the training set for the machine learning algorithm.

Table S3 AGI codes and accession numbers of validated plant alternative proteins.

Table S4 The full list of analyzed isoform pairs from animals, in the order corresponding to the graph presented in Fig. 3(b).

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.