ASES: visualizing evolutionary conservation of alternative splicing in proteins

Diego Javier Zea, Hugues Richard, Elodie Laine

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Sequence analysis

ASES: Visualising evolutionary conservation of alternative splicing in proteins

Diego Javier Zea [1], Hugues Richard [2,*] and Elodie Laine [1,*]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.
[2] Bioinformatics Unit (MF1), Department for Methods development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany

* To whom correspondence should be addressed: RichardH@rki.de, elodie.laine@sorbonne-universite.fr

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: ASES is a versatile tool for assessing the impact of alternative splicing, initiation and termination of transcription on protein diversity in evolution. It identifies exon and transcript orthogroups from a set of input genes/species for comparative transcriptomics analyses. It computes an evolutionary splicing graph, where the nodes are exon orthogroups, allowing for a direct evaluation of alternative splicing conservation. It also reconstructs a transcripts' phylogenetic forest to date the appearance of specific transcripts and explore the events that have shaped them. ASES web server features a highly interactive interface enabling the synchronous selection of events, exons or transcripts in the different outputs, and the visualisation and retrieval of the corresponding amino acid sequences, for subsequent 3D structure prediction.
Availability and Implementation: http://www.lcqb.upmc.fr/Ases.
Contact: RichardH@rki.de, elodie.laine@sorbonne-universite.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative splicing, alternative promoter usage or alternative polyadenylation have the potential to greatly expand proteomes by generating multiple mRNA transcripts from a single gene. These transcripts may lead to protein isoforms with different shapes and functions. Alternative splicing (AS) has been linked to morphological diversity (Tovar-Corona *et al.*, 2015), organ development (Mazin *et al.*, 2021), disease susceptibility (Park *et al.*, 2018), immune adaptation (Rotival *et al.*, 2019) and interactome rewiring (Yang *et al.*, 2016), among others. Hence, there is a growing interest in assessing its functional impact at the protein level, and in determining the evolutionary origin of alternative protein isoforms. To this aim, we have recently developed a couple of efficient scalable computational methods. The first one, ThorAxe (Zea *et al.*, 2021), adds an evolutionary dimension to the analysis of transcript variability by extending the notion of splicing graph to an ensemble of genes/species. It establishes a mapping between orthologous regions in a set of transcripts coming from different genes/species, and uses this mapping to decompose the transcripts into building blocks. The second one, Phylosofs (Ait-Hamlat *et al.*, 2020), starts from ThorAxe's description of transcripts and takes a step further by inferring plausible scenarios explaining transcripts' evolution. Both tools adopt an end-product perspective where the transcripts are represented by their translated amino acid sequences. This choice allows reasoning directly over the impact of AS-induced variations on the protein sequences and structures. Here, we present Alternative Splicing Evolution Server (ASES), a versatile tool to explore the contributions of AS to protein diversity in evolution. Starting from a set of transcripts annotated in Ensembl (Zerbino *et al.*, 2018), it relies on ThorAxe and PhyloSofS to compute an evolutionary splicing graph (ESG) and a transcripts' phylogenetic forest (TPF). ASES allows configurable queries enabling users to interactively inspect results at all scales (residue, exon, AS-event, transcript). It provides a rich environment for identifying the mechanisms underlying protein functional diversification through AS.

2 Definitions and implementation

The ESG is a compact exon-centred representation of the transcript variability observed in a set of genes/species (Figure 1A). The nodes represent s-exons, *i.e.* minimal transcript building blocks defined across species. The edges represent consecutive co-occurrences of s-exons in the input transcripts. Hence, a transcript is a path in the ESG. We estimate the

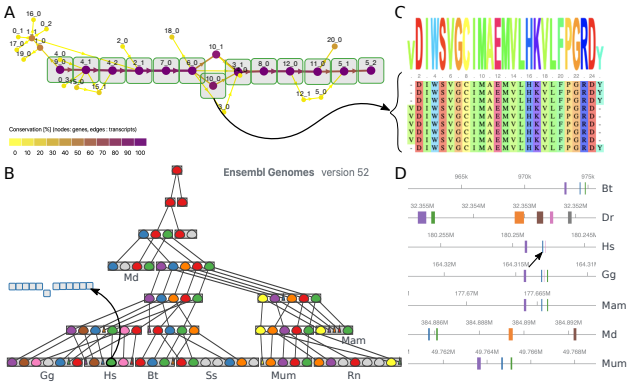

Fig. 1. ASES example outputs for the human gene *MAPK9* and its orthologs in 9 species. **A**. ESG, with colors indicating conservation levels (node: proportion of species, edges: average proportion of transcripts). The grey rectangles highlight the path corresponding to the green transcript selected in panel B. **B**. TPF, where the nodes (transcripts) with the same colour belong to the same tree. The grey leaves are orphan transcripts (no phylogeny). The set of leaves observed in a present-day gene/species are grouped in a grey rectangular box. The boxes are transparent for ancestral genes/species.The triangles indicate transcript losses. **C**. Consensus sequence and MSA of the s-exon *10_0* selected in panel A. The human sequence is on the first row. **D**. Close-up view of the gene structure, where each line corresponds to a gene/species and each color indicates a s-exon. The blue box pointed by the arrow is the human exonic region belonging to the s-exon *10_0* selected in panel A.

evolutionary conservation of a s-exon (node) or a s-exon junction (edge) by quantifying the amount of species and transcripts including it, respectively. The heuristic for computing the ESG unfolds in five main steps (Zea *et al.*, 2021): (1) cluster similar exons together, (2) define sub-exons within each species, (3) create a multiple sequence alignment (MSA) for each cluster, (4) identify s-exons as continuous blocks in the MSA, (5) refine s-exons. The TPF is a transcript-centred representation of the evolution of AS (Figure 1B). Each tree gives a plausible scenario for the emergence of a new transcript in evolution (the root) and how it led to a set of transcripts observed in present-day genes/species (the leaves). The whole forest is embedded in the gene tree retrieved from Ensembl. We rely on Sankoff's algorithm to determine the most parsimonious TPF (Ait-Hamlat *et al.*, 2020). We implemented ASES web server using Genie, a Julia web framework (Bezanson *et al.*, 2017), with JavaScript and JQuery for front-end interactivity. We used Cytoscape.js (Franz *et al.*, 2016), neXtProt viewer and BioJS MSAViewer (Yachdav *et al.*, 2016) to visualise the ESGs, the protein sequences and the MSAs, respectively. D3 JavaScript library (Bostock *et al.*, 2011) is used for the forest and the gene structure.

3 Usage

ASES accepts as input the name or the Ensembl (Zerbino *et al.*, 2018) stable identifier of a gene, the species where it comes from, and a list of species for comparative analysis – human (Hs), gorilla (Gg), macaque (Mam), mouse (Mum), rat (Rn), boar (Ss), cow (Bt), opossum (Md), platypus (Oa), frog (Xt), zebrafish (Dr) and nematode (Ce) by default. Once the job is submitted, the computation is done in three steps. First, ThorAxe dynamically retrieves the annotations associated to the query gene (genomic exons, protein coding transcripts, gene tree...etc) and its orthologs in the considered species from Ensembl. By default, we consider only the one-to-one orthologs defined in Ensembl, and we exclude the species containing multiple orthologs from the analysis. Optionally, users can extend the analysis to multiple orthology relationships. Then, ThorAxe computes the ESG and prepares the input for Phylosofs, which in turn infers the TPF. This last step typically takes more time and can be turned off.

ASES generates interactive representations of the ESG and the TPF, an interactive map of the gene structure with explicit genomic coordinates, and two tables (Figure 1). The colouring is intended to ease and guide users navigation. For instance, in the ESG, the nodes and edges get darker with their conservation degree (Figure 1A). Hence, although the ESG may seem complex at first sight, many AS events (bubbles in the graph) are typically not conserved (in yellow). See Supplementary Table S1 for statistics at the human protein-coding genome scale. In *MAPK9*, we readily identify a conserved mutually exclusive event involving the s-exons *10_0* and *10_1*. ASES dynamically links the ESG and the TPF, thereby facilitating the inspection of specific s-exons, events or transcripts. Here, the path highlighted in the ESG (Figure 1A) corresponds to the human green transcript selected in the TPF (Figure 1B). It includes the s-exon *10_0* and originates from an ancestral transcript (green root) present in the ancestor common to eutheria (Gg, Hs, Bt, Ss, Mum, Rn, Mam) and metatheria (Md). Among the five transcripts predicted for the eutheria ancestor, three were lost in human (purple, blue and red triangles). The users can visualise the s-exon losses and gains along the phylogeny upon clicking on the branches. ASES also displays the protein isoform sequences, and the user may simply copy-paste them in a domain-annotation or 3D-structure prediction web server. Going back to the ESG (Figure 1A), selecting one s-exon, *10_0* here, makes the associated MSA appear (Figure 1C). We can see that *10_0* has a 25 AA consensus and is almost perfectly conserved across 9 species. Users can also select s-exons in the genomic map (Figure 1D). Here, we see that the s-exon *10_0* (see arrow) is absent from zebrafish. This observation is reflected by the TPF, where the red transcripts include *10_1* instead of *10_0*. Finally, the tables give detailed information about each of the s-exons and events, *e.g.* genomic coordinates, conservation levels, list of species, length and sequences. All data can be downloaded for a local analysis. ASES can be used in complement to other ressources, for instance to investigate the evolutionary origin and per-s-exon conservation of transcripts annotated in APPRIS (Rodriguez *et al.*, 2013) (see Supplementary Figure S1).

Funding

A grant of the French national research agency (MASSIV project, ANR-17-CE12-0009) provided a salary to D.J.Z.

References

Ait-Hamlat, A. *et al.* (2020). Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the jnk family. *Journal of molecular biology*, **432**(7), 2121–2140.

Bezanson, J. *et al.* (2017). Julia: A fresh approach to numerical computing. *SIAM review*, **59**(1), 65–98.

Bostock, M. *et al.* (2011). D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, **17**(12), 2301–2309.

Franz, M. *et al.* (2016). Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**(2), 309–311.

Mazin, P. V. *et al.* (2021). Alternative splicing during mammalian organ development. *Nature Genetics*, **53**(6), 925–934.

Park, E. *et al.* (2018). The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics*, **102**(1), 11–26.

Rotival, M. *et al.* (2019). Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nature communications*, **41**(Database issue), 1–15.

Rodriguez13, J. M. *et al.* (2013). APPRIS: annotation of principal and alternative splice isoforms *Nucleic Acids Research*, **10**(1), D110–117.

Tovar-Corona, J. M. *et al.* (2015). Alternative splice in alternative lice. *Molecular biology and evolution*, **32**(10), 2749–2759.

Yachdav, G. *et al.* (2016). Msaviewer: interactive javascript visualization of multiple sequence alignments. *Bioinformatics*, **32**(22), 3501–3503.

Yang, X. *et al.* (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**(4), 805–817.

Zea, D. J. *et al.* (2021). Assessing conservation of alternative splicing with evolutionary splicing graphs. *Genome Research*, pages gr–274696.

Zerbino, D. R. *et al.* (2018). Ensembl 2018. *Nucleic Acids Res.*, **46**(D1), D754–D761.