IsoSel: Protein Isoform Selector for phylogenetic reconstructions

H. Philippon, A. Souvane, Céline Brochier-Armanet, G. Perriere

PLOS ONE

RESEARCH ARTICLE

IsoSel: Protein Isoform Selector for phylogenetic reconstructions

Héloïse Philippon, Alexia Souvane, Céline Brochier-Armanet, Guy Perrière*

Univ. Lyon, Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, F-69622, Villeurbanne, France

* guy.perriere@univ-lyon1.fr

Citation: Philippon H, Souvane A, Brochier-Armanet C, Perrière G (2017) IsoSel: Protein Isoform Selector for phylogenetic reconstructions. PLoS ONE 12(3): e0174250. https://doi.org/10.1371/journal.pone.0174250

Editor: Olivier Lespinet, Universite Paris-Sud, FRANCE

Received: October 21, 2016

Accepted: March 6, 2017

Published: March 21, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Sofware described in the manuscript can be freely downloaded at http://doua.prabi.fr/software/isosel.

Funding: This work was supported by grants from the CNRS, the Institut Français de Bioinformatique to GP, the France Génomique consortium for GP, and the Région Rhône-Alpes PhD thesis grant 13-012735-01, which was funding the Ph.D of HP. CBA is a member of the Institut Universitaire de France and is funded by the "Ancestrome" project

Abstract

The reliability of molecular phylogenies is strongly dependent on the quality of the assembled datasets. In the case of eukaryotes, the selection of only one protein isoform per genomic locus is mandatory to avoid biases linked to redundancy. Here, we present IsoSel, a tool devoted to the selection of alternative isoforms in the context of phylogenetic reconstruction. It provides a better alternative to the widely used approach consisting in the selection of the longest isoforms and it performs better than Guidance, its only available counterpart. IsoSel is publicly available at http://doua.prabi.fr/software/isosel.

Introduction

The alternative splicing, a process by which a single coding gene may lead to different transcripts and thus to different protein isoforms, is common in Eukaryotes. For instance, about 20% of plant genes [1] and 90% of human genes [2] undergo alternative splicing. In molecular phylogeny, the construction of homologous sequences datasets—usually performed by similarity-based procedures—does not allow distinguishing among the various isoforms and all of them are gathered during the process. However, most of the time only one isoform is kept for phylogenetic analyses, because they carry redundant information. Furthermore, due to the fact that some exons are present in some isoforms and absent in others, aligning them frequently leads to the introduction of many gaps in Multiple Sequence Alignments (MSAs) [3]. Finally, trimming programs like Gblocks [4] or BMGE [5] select alignment regions based on their conservation level. So, introducing many isoforms will lead to the overestimation of conservation rates, a same residue being artefactually represented many times in the MSA.

It is therefore absolutely necessary to select a unique sequence per genomic locus before reconstructing a phylogeny. Although manual selection provides usually the best results, this becomes tedious when the number of sequences and/or of homologous families is large. Two simple automated approaches are therefore commonly used: the random selection of one isoform [6, 7] or the selection of the longest isoform [8, 9]. But there is no conceptual justification for the former and the latter usually leads to the introduction of many gaps in the alignments. Presently, only one software dedicated to isoform selection is available: PALO [10]. It selects the combination of isoforms that are most similar in length. However, PALO only works on

through the Agence Nationale de la Recherche grant ANR-10-BINF-01-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

sequences from the Ensembl database and its unique selection criterion is based on sequence length.

In that context, we developed IsoSel (Isoform Selector), a tool designed to the selection of protein isoforms specifically designed for phylogenetic reconstruction. IsoSel is based on the same approach as the one used by Guidance [11], a tool devoted to the assessment of protein MSAs reliability. The two main differences between IsoSel and Guidance are the availability of a broader range of amino acids substitution models in IsoSel and the introduction of different options allowing to penalize short and long sequences. IsoSel can also provide an output file giving the corresponding selected isoform for each genomic locus. Most frequently, the sequences selected by IsoSel allows to build phylogenetic trees that are better than those obtained with the sequences selected by Guidance, this when considering a tree length criterion and the number of Duplication-Loss (DL) events inferred in a tree reconciliation. Lastly, IsoSel is a standalone program that does not require the availability of interpreted languages such as Perl and Ruby. It is therefore easier to install and to use.

Materials and methods

Algorithm

The first step of an IsoSel run consists in the alignment of an input protein dataset, using either CLUSTALO [12], MAFFT [13] or MUSCLE [14], to generate what we call a reference alignment. We selected those three programs because they allow to input a user-provided guide tree when building a MSA, which is required during the second step of the algorithm (Fig 1).

The second step is the generation of a set of perturbed alignments through a bootstrap approach. Let a be the reference alignment obtained during step 1, ℓ the length of this alignment and n the number of bootstrap replicates set by the user. First, n alignment replicates are generated by the standard bootstrap procedure (*i.e.*, random sampling with replacement of ℓ sites among a). For each bootstrap replicate, a distance matrix is then computed by IsoSel, this using one of the amino acid substitution models implemented in the program: Poisson (with or without Gamma correction), PAM (or its Kimura approximation), JTT (or its Gamma-corrected Poisson approximation), BLOSUM62, WAG or LG. From this distance matrix, the BioNJ [15] algorithm is used to infer a tree. Finally, the input dataset is realigned using this tree as guide tree, this with the same MSA program as the one used the first step. The n resulting realignments of a represent the perturbed alignments.

The third and final step is the computation of the Sum-of-Pairs (SP) score [16] using the perturbed alignments. As described in the original publication, the SP score is used as a comparison metric between two MSAs (Fig 2).

Let m be the number of sequences of the input dataset and b ($1 \leq b \leq n$) one of the n perturbed alignments generated during the second step. For each sequence i ($1 \leq i \leq m$) of length l_i from b, the pair score of the amino acid at position k ($1 \leq k \leq l_i$) is computed as:

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m_k} p_{ijk} \quad (1)$$

where p_{ijk} is equal to 1 or 0 whether or not the amino acid at position k of sequence i is facing the same amino acid at the position k' ($1 \leq k' \leq l_j$) of sequence j in the reference and perturbed alignment (Fig 2). Also, m_k is the number of sequences having a residue (*i.e.*, not a gap) at this position in a. If there no other residue at this position (*i.e.*, if $m_k = 1$), $R_{ik}^{(b)}$ is set to 1.

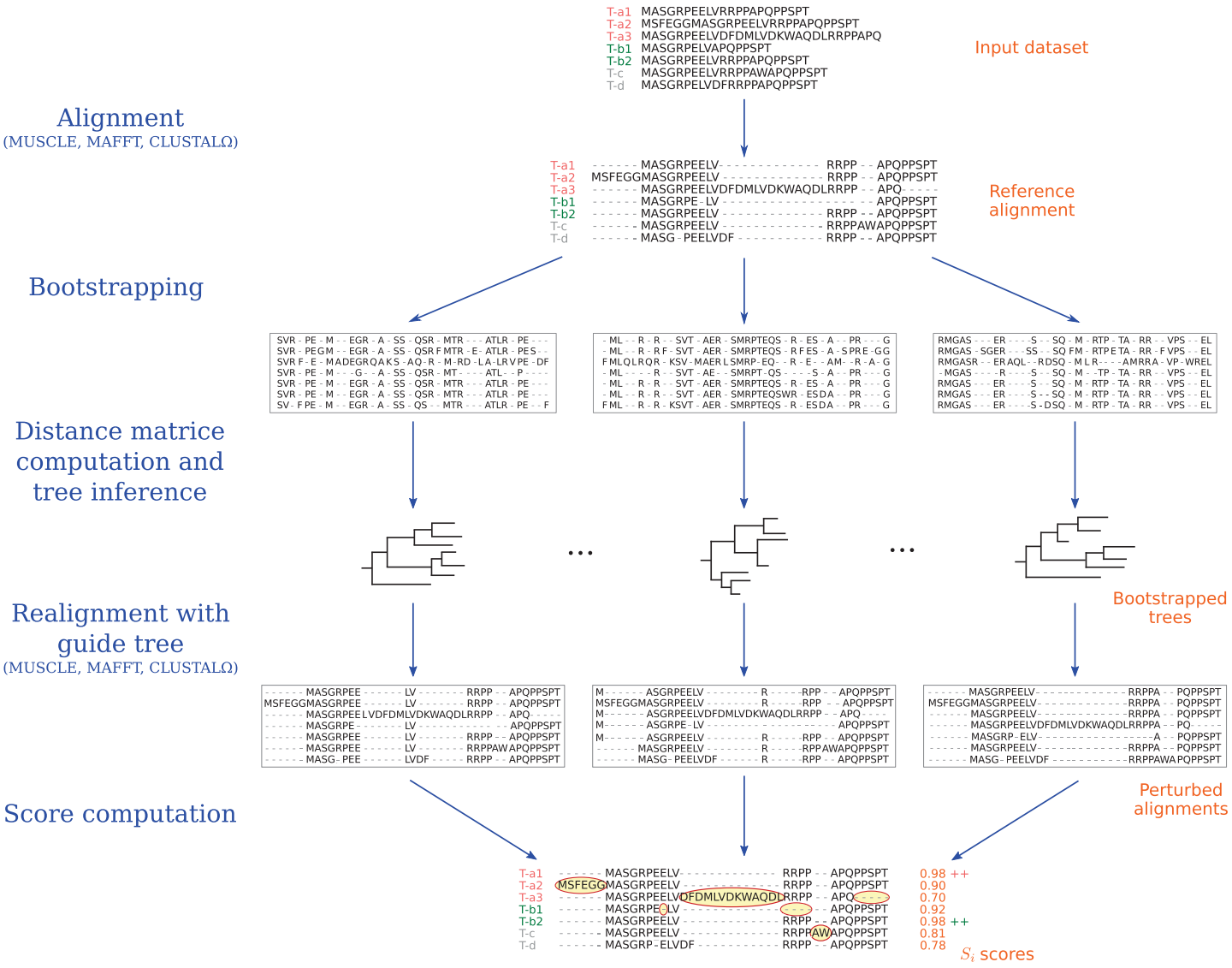

Fig 1. IsoSel workflow. Schematic representation of the different steps performed during an IsoSel run. T-x represent alternatives isoforms generated by a same gene x. In this example, isoforms a1 and b2 are selected for the genes a and b, respectively.

https://doi.org/10.1371/journal.pone.0174250.g001

From Eq (1) it is possible to compute the average residue score over all bootstrap replicates as:

$$R_{ik} = \frac{1}{n}\sum_{b=1}^{n} R_{ik}^{(b)} \tag{2}$$

Finally, the SP score S_i for sequence i is calculated by averaging the residues scores:

$$S_i = \frac{1}{l_i}\sum_{k=1}^{l_i} R_{ik} \tag{3}$$

By construction, $0 \leq S_i \leq 1$. The higher its value, the better the alignment of the isoform with the other sequences. For each genomic locus, the isoform with the best score is kept.

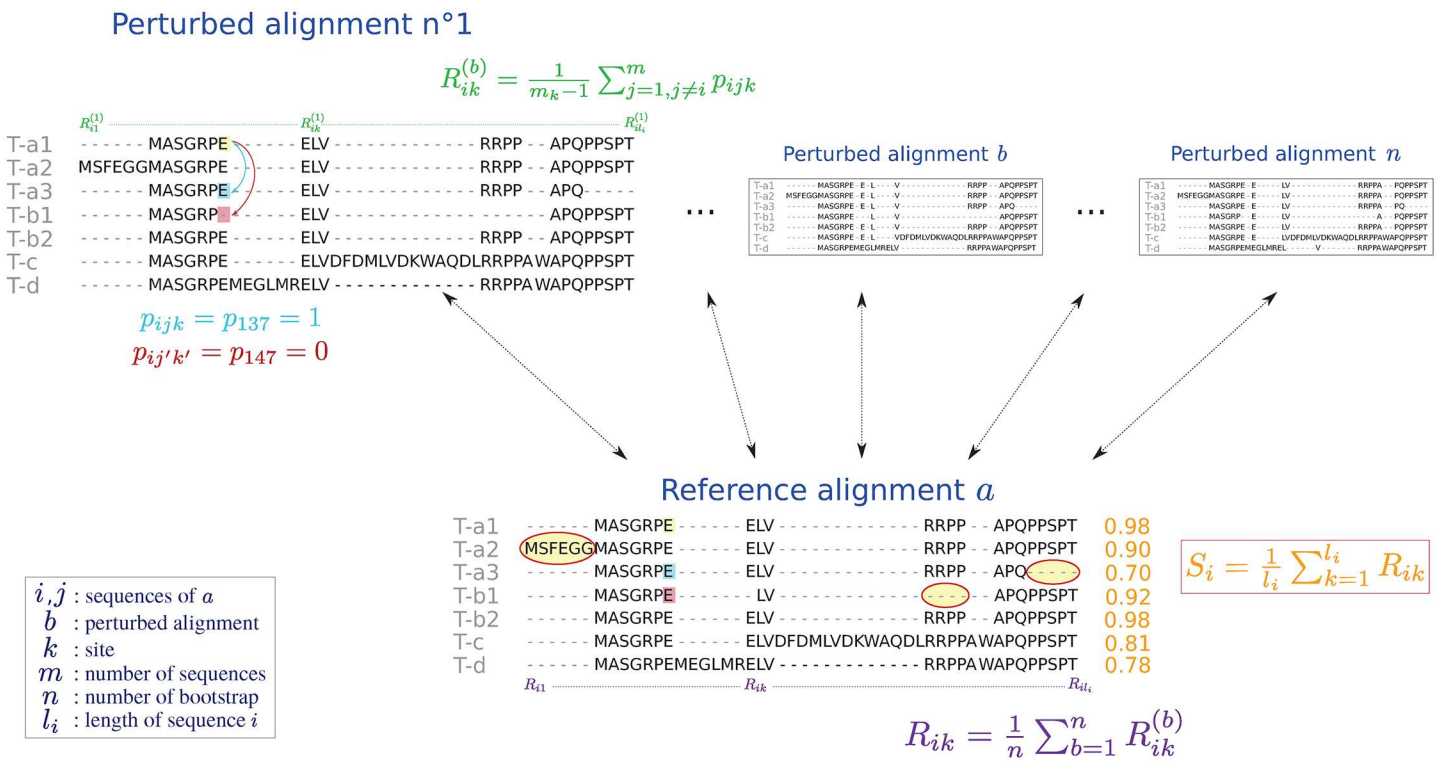

Fig 2. SP score computation. Example of score computation for four genes (a, b, c and d) producing three (a), two (b) and no (c and d) alternative isoforms.

https://doi.org/10.1371/journal.pone.0174250.g002

Special options

IsoSel comes with a lot of available options and parameters (for the substitution models, the Gamma correction, the alignment programs, etc.) and we will not list them all. But some of those options are of special interest as they can allow to improve greatly the results depending on the dataset contents. Below is a short description of those key options and in which context they should be used.

The `-gap` option. The standard SP score described in the previous section does not take into account the number of gaps present at a given site of a. For that purpose, we have implemented an option (`-gap`) which penalizes the sequences introducing gaps in the reference alignment. With this option, p_{ijk} is weighted by the number of gaps present at position k in a:

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m_k} \frac{p_{ijk}}{m_g} \quad (4)$$

where $m_g = m - m_k$ is the number of gaps at this position in a. If there is no gap, then $m_g = 1$. Here, p_{ijk} is the same as before excepted that it is fixed at −1 if there is no other residue at this position in the reference alignment. By construction, S_i values are computed using the same equation as before but they are much smaller. Nevertheless the alternative transcript selected is always the one with the highest score. This option is suited for datasets containing alternative isoforms resulting from intron retention or exon skipping events that are restricted to a reduced set of taxonomic groups.

The `-short` option. If the option `-gap` penalizes too long isoforms, the option `-short` was designed to penalize short isoforms, even if they are well aligned. In this case, S_i is no

longer calculated by dividing the sum of residue scores by the length of the sequence, but by the length of the reference alignment:

$$S_i = \frac{1}{\ell} \sum_{k=1}^{l_i} R_{ik} \quad (5)$$

As for the `-gap` option, the S_i values obtained are, by construction, smaller. This option is suited for datasets with isoforms corresponding to partial sequences or when a very short isoform is specific to a given taxonomic group.

The `-DS` and `-WOT` options. The bootstrap approach, with the computation of a set of perturbed alignments, is very time consuming when the number and/or the length of the sequences increase. To allow a faster approach, the `-DS` option (for "Distance Scores") approximates the SP score by the mean of observed divergence (*p*-distance) between sequence *i* and the other input dataset sequences. Unlike the other scoring schemes, there is no bootstrap resampling procedure and the isoform selected is the one with the smallest distance (*i.e.*, the sequence selected is the one that it is the more similar to the others).

In order to take into account the difference of lengths between isoforms, the `-DS` option uses a modified *p*-distance in which a gap is considered as a supplementary character state. In that way, considering two isoforms that only differs by an exon skipping event, their modified *p*-distance will not be near to one as it is with the standard *p*-distance. A bias towards the systematic selection of smaller isoform is thus avoided.

Let d_{ij} be the modified *p*-distance between sequences *i* and *j* ($1 \leq i, j \leq m$). The score S_i corresponding to the sequence *i* is then computed as:

$$S_i = \frac{1}{m_s} \sum_{i \in \Omega, i \neq j} d_{ij} \quad (6)$$

where Ω corresponds to the subset of sequences without isoform and m_s its cardinal. The higher S_i, the more distant to the other sequences of the input dataset is the sequence *i*.

In combination to the `-DS` option, the `-WOT` option (for "With Other Transcripts") computes the mean of distances on the entire input dataset. Then Eq (6) becomes:

$$S_i = \frac{1}{m_s} \sum_{i=1, i \neq j}^{m} d_{ij} \quad (7)$$

This option is suited for datasets containing a majority of homologous genes generating alternative isoforms.

The `-auto` option. Best options for IsoSel are dataset dependent, therefore we have implemented an automated mode (`-auto`) allowing to estimate them. Under this mode, if the reference alignment has $>$ 35% of sites containing $>$ 80% gaps, the S_i scores are computed with the `-gap` option. If there are $>$ 600 sequences in the input dataset or if the alignment length is $>$ 10000 AA, then the option `-DS` is selected. In all other cases, the default parameters are used.

Input and output

IsoSel minimal input requirement is an unaligned set of protein sequences in Fasta format. The output is a text file containing the scores for each input sequence (Fig 3). Optionally, the user can provide a file in which the information on transcripts locus tag is given. In this case, IsoSel will also create a file in Fasta format that will contain the filtered dataset (*i.e.*, in which only the isoform having the best score for a given gene is kept).

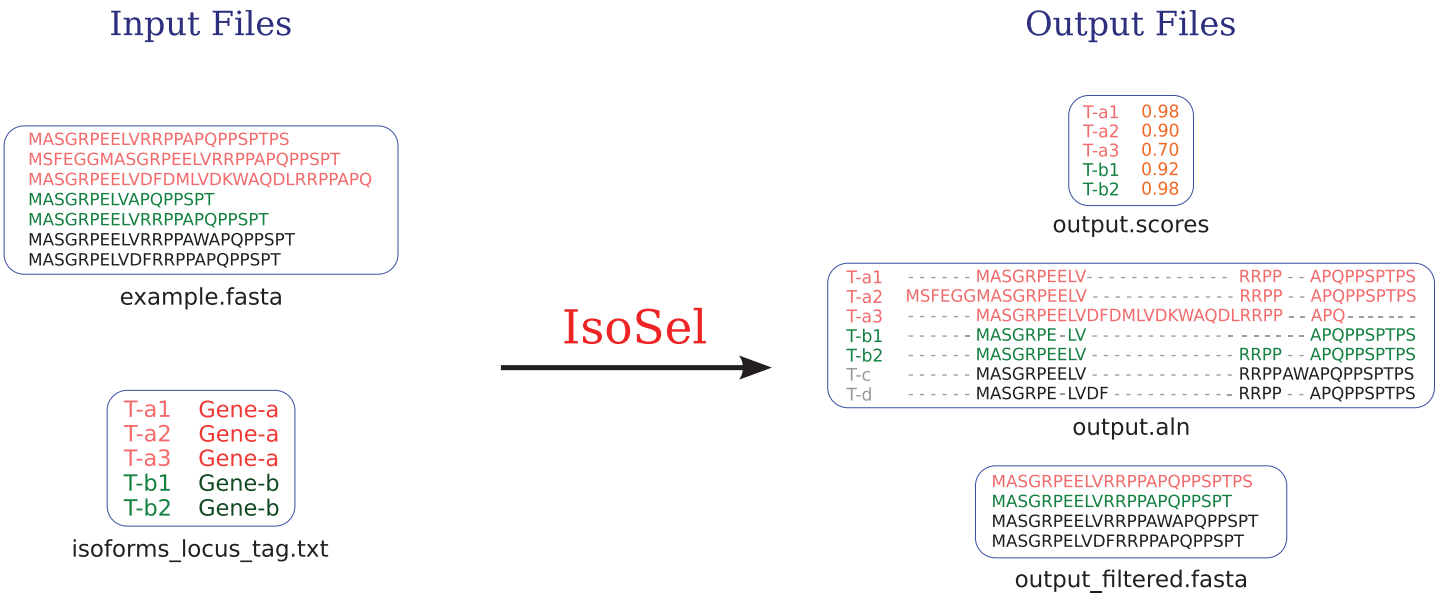

Fig 3. Input and output files. IsoSel minimal input requirement is an unaligned protein sequence dataset in Fasta format (`example.fasta`). The two output files generated contain the alignment (`output.aln`) and the sequences scores (`output.scores` or `output.DistanceScore` if the `-DS` option is used). Optionally, the user can provide a file containing the genomic origin of the input sequences (`isoforms_locus_tag.txt`). In this case, an additional file containing, for each locus, the sequence having the highest score is created (`output_filtered.fasta`).

https://doi.org/10.1371/journal.pone.0174250.g003

An example dataset containing two files is included in the program distribution. The first one, named `example.fasta`, contains 58 homologs of the human AKTS1 protein taken from Ensembl and from a local database of complete eukaryotic proteomes. Among those 59 sequences, 26 correspond to alternative isoforms whose genomic origin is indicated in the second file, named `isoforms_locus_tag.txt`. The first column of this second file corresponds to the sequence names (the same as those in the `example.fasta` file) and the second column to an identifier allowing associating a set of isoforms to a given gene. For example, the three lines below:

```
>ENSP00000375710|Homo_sapiens ENSG00000204673
>ENSP00000375711|Homo_sapiens ENSG00000204673
>ENSP00000375706|Homo_sapiens ENSG00000204673
```

allows to specify that sequences ENSP00000375710, ENSP00000375711 and ENSP00000375706 are isoforms that originate from a single gene, identified as ENSG00000204673.

Datasets for program testing

We randomly sampled 200 human proteins among the 20201 available in UniProtKB release 2016_05. For each sequence, we searched for its homologs into two collections using BLASTP [17], this with a similarity threshold set at $E \leq 10^{-30}$. The first collection corresponded to a subset of Ensembl release 80 containing 32 species while the second was made of 84 complete eukaryotic proteomes taken from GenBank release 70 (S1 and S2 Tables, respectively). Among the 200 human proteins used as seed for BLASTP, 12 corresponded to ORFans and 32 returned less than 20 homologous genes. Moreover, two BLASTP runs contained only one

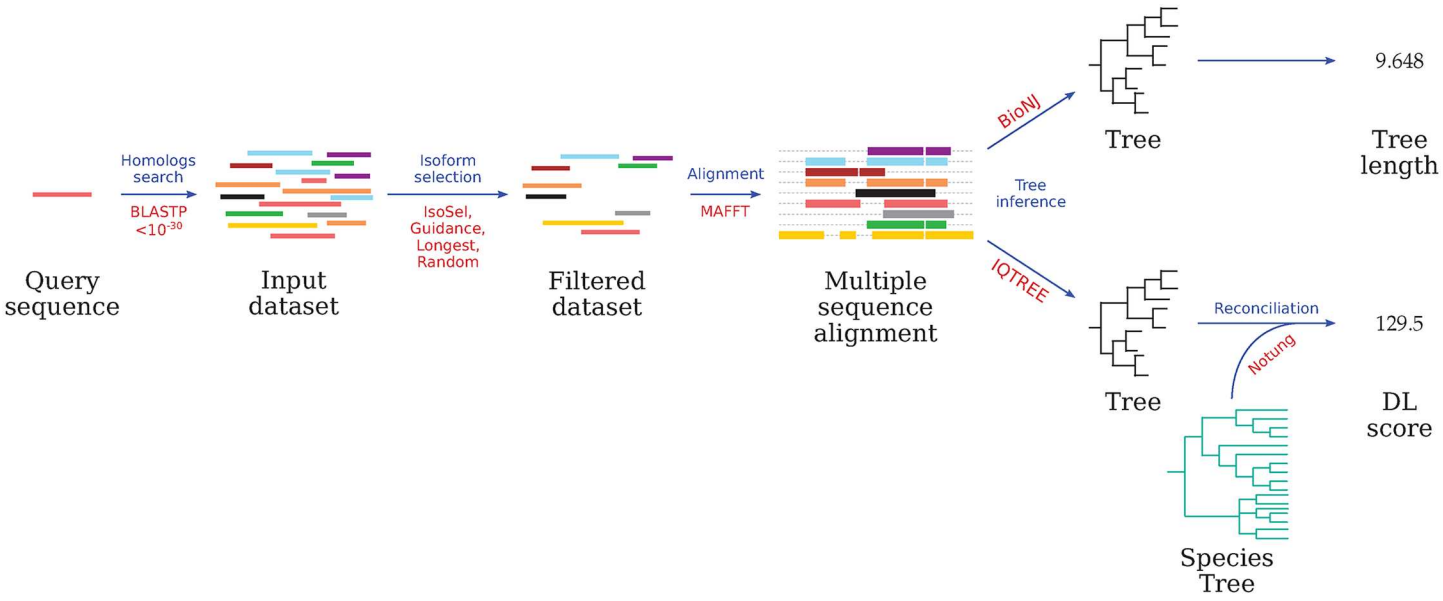

Fig 4. Workflow used for testing IsoSel performances. For a given human protein from UniProtKB, a BLASTP search is performed. The alternative isoforms detected for each set of homologs are then selected using either the longest isoform, a random choice, Guidance or IsoSel. Then the sets are aligned and the corresponding gene trees are inferred by BioNJ and IQ-TREE for computing tree lengths and DL scores, respectively. For each step, algorithms used are indicated in red.

https://doi.org/10.1371/journal.pone.0174250.g004

gene with two alternative isoforms. All those sets were discarded and the remaining 154 were used to test the performances of IsoSel relatively to the other possible strategies (S3 Table).

The complete workflow summarizing the testing procedure is shown in Fig 4. For each of the 154 sets of homologs, we ran IsoSel with the different available options and 30 bootstraps replicates. We compared the results returned by IsoSel to selections obtained with: i) the longest isoforms; ii) the random choice of an isoform for each genomic locus; and iii) the scores obtained using Guidance 2.0 [18] with 30 bootstrap replicates and default parameters.

The resulting filtered datasets, containing only one sequence per genomic locus, were aligned using MAFFT (automatic algorithm selection and a maximum of 20 iterations). From each MSA, two trees were inferred to compute the tree length and the DL score. The first tree was inferred by SeaView [19] using the *p*-distance and the BioNJ algorithm. Then, the sum of the branch lengths was computed and all tree lengths obtained are listed in S4 Table. For the second gene tree inference, we run BMGE to select conserved blocks of the MSA (BLOSUM30 substitution matrix, 40% of gap allowed and a minimal block length of three amino acids). Then, the selection of evolutionary models (Bayesian Information Criterion) and the tree inference was computed by maximum likelihood using IQ-TREE [20] with default parameters and 1000 replicates for the Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT). Finally, the resulting gene tree was rooted and reconciled with a reference tree of eukaryotic species [21, 22] (S1 Fig) using Notung [23]. All the DL scores obtained are listed in S5 Table.

Among the 154 sets, we used the one built with the human protein WDR18 (UniProtKB accession number Q9BV38) as a case study. For this sequence, the BLASTP search in the two collections led to a set of 73 homologous protein sequences. Among them, fourteen (19.2%) resulted from alternative splicing events (S3 Table). Selection of isoforms from this set was performed by the longest sequence criterion and by IsoSel (default parameters). For both

selections (containing 63 sequences) we applied the methodology described above for computing maximum likelihood trees. The two filtered alignments contained a total of 349 and 354 conserved sites, respectively. In both cases, the evolutionary model selected by IQ-TREE was LG+F+Γ_4 [24]. The resulting trees were then formatted using TreeGraph2 [25].

Results

Tree length criterion

There is presently no gold standard to evaluate the quality of a MSA relatively to a phylogenetic criterion. However, Rzhetsky and Nei [26], stated that the tree with the smallest sum of branches length is most likely to be the true one. We therefore used this criterion to roughly estimate the quality of the MSAs. We compared the length of the trees generated with the IsoSel filtered datasets to those obtained by keeping a random isoform, the longest isoform or by a selection with Guidance. For that purpose, we used the 154 sets of eukaryotic protein families obtained through the procedure described above. For more than 76% of the test datasets, the use of IsoSel led to a shorter tree than the one obtained using the other approaches (S4 Table and Fig 5A).

The longest and the random isoforms selections led to the smallest trees in only three and five cases, respectively. Moreover, for 129 and 127 of the 154 datasets (corresponding to 83.7% and 82.4% of the total), using IsoSel with the `-auto` option led to shorter trees than the random and the longest isoform selections, respectively. A Wilcoxon paired test showed that these results are highly significant (both $P < 2.2 \times 10^{-16}$). In 119 cases (78.8%), one of the IsoSel options led to a shorter tree than the Guidance selection, which is also highly significant ($P < 5.27 \times 10^{-12}$).

DL score criterion

Another criterion of gene tree quality is its congruence to a reference species tree, especially in the case of eukaryotic species where horizontal gene transfers are rare events. We therefore performed a tree reconciliation for each of our datasets using Notung. The DL score provided by this program is proportional to the minimum number of duplications and losses needed to reconcile a gene tree with a species tree. Therefore, the lower the score is, the closer the gene tree is to the species tree.

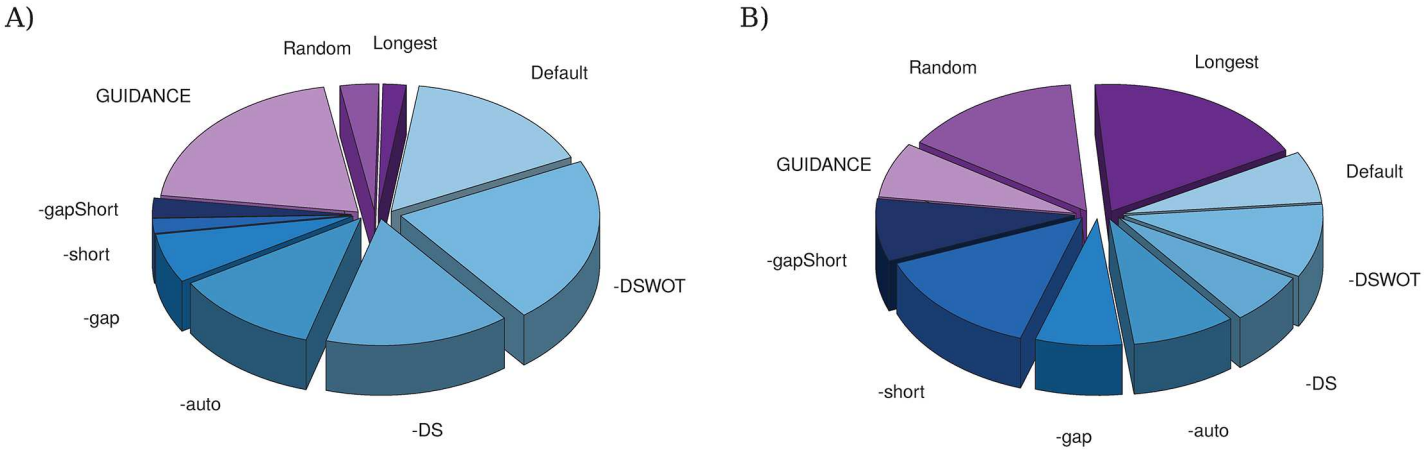

Fig 5. Tree lengths and DL scores distributions. Charts are proportional to the number of: A) the shortest trees; and B) the trees with the lower DL score obtained with the different options and programs. Charts in shades of blue correspond to the different IsoSel options.

https://doi.org/10.1371/journal.pone.0174250.g005

For 122 datasets (corresponding to 79.2% of the total), one of the IsoSel options allowed to obtain a tree with an equal (20 cases) or better (102 cases) DL score than the one inferred after the selection of the longest isoform. The comparison with the random selection gave similar results as IsoSel was better in 117 cases and equal in seven (corresponding to 75.9% and 4.5%, respectively). Finally, in 128 cases (84.7%), IsoSel gave better results than the Guidance selection and performed equally in thirteen cases (8.6%). All those results are highly significants (all Wilcoxon paired tests led to $P < 10^{-8}$). Globally, for 63.5% of the datasets, the use of IsoSel led to equal or less discordant gene trees than those obtained using the other approaches (Fig 5B).

WDR18 protein

The phylogenetic trees obtained using with the longest and the IsoSel (with default parameters) procedures are shown in Fig 6. The selection carried out by IsoSel led to a shorter but more discordant gene tree than the selection using the longest isoforms (S4 and S5 Tables). The selection using the longest sequences led to a phylogenetic tree in which the isoform selected for *Monodelphis domestica* (ENSMODT00000007188) is misplaced, this erroneous placement being probably linked to the long branch generated (Fig 6A). *Saccoglossus kowalevskii* is also incorrectly placed in the selection using the longest sequences. Its position is improved in the IsoSel selection but there is no statistical supported for the placement in both trees. Another consequence of the longest isoforms selection is that the clade grouping mammals other than *M. domestica* is not supported, while it is in the tree obtained with the IsoSel

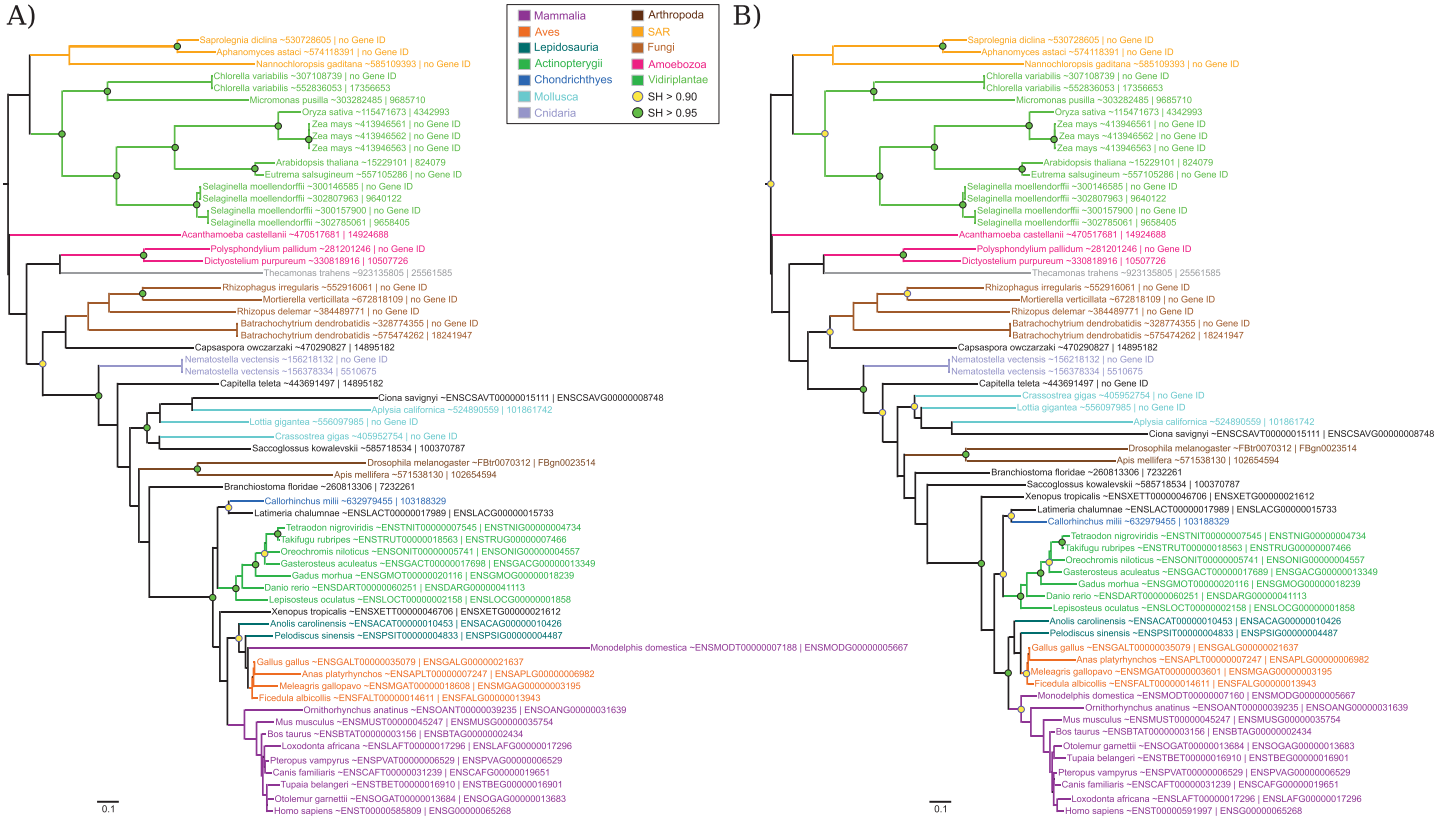

Fig 6. Maximum likelihood trees for WDR18 protein. Isoform selection was done by selecting the longest isoform (A) and by running IsoSel with its default parameters (B). Sequences are colored according to their taxonomic classification. Green and yellow circles correspond to nodes with SH > 0.95 and SH > 0.90, respectively. The scale bar represents the average number of substitutions per site.

https://doi.org/10.1371/journal.pone.0174250.g006

set (SH = 0.92, Fig 6B). Similarly, the clade corresponding to the Amniotes is only supported in the tree obtained with the IsoSel selection (SH = 0.98). On the other hand, according to our reference tree for eukaryotic species (S1 Fig), the human sequence is incorrectly grouped with the elephant and *Xenopus laevis* is misplaced in the tree built with the IsoSel selection.

Conclusion

IsoSel is a command line software designed for the automatic selection of protein isoforms in the framework of phylogenetic analyses. Based on the SP score, it allows to obtain datasets that are optimized for tree reconstruction. The only other software that can be compared to IsoSel is Guidance but this program presents some limitations. First, it requires the independent installation of a broad range of tools (namely Perl, BioPerl and Ruby) while IsoSel is self-sufficient and is distributed with all the binaries required for its functioning. Then, it can only be run with the JTT substitution model while IsoSel allows the use of all standard site-homogeneous models. On a practical point of view, Guidance is usually slower than IsoSel when multithreading is enabled (data not shown). This point is probably linked to the fact that Guidance was not designed for alternative isoforms selection but rather as a general tool for assessing MSA quality. With this broader purpose, Guidance has to compute many scores in addition to SP, which lower its performances in terms of speed.

Globally, it appears that, compared to the other available approaches, IsoSel allows selecting most frequently the sequences leading to gene trees that are shorter and closer to the species tree. It is thus more suited for phylogenetic reconstructions. IsoSel is implemented in C/C++, is optimized for multithreading and is available under the CeCILL license.

Supporting information

S1 Table. Sizes of the proteomes selected from Ensembl. Number of protein sequences available in Ensembl for each of the 32 selected species.
(PDF)

S2 Table. Sizes of the complete eukaryotic proteomes selected from GenBank. Number of protein sequences for each selected eukaryotic complete genomes. The selected species are different from the ones from Ensembl.
(PDF)

S3 Table. Characteristics of the 200 test datasets. For each of the randomly selected human protein, the number of homologous sequences detected using BLASTP is indicated in the third column. The fourth and fifth columns give the gene number and percentage of alternative isoforms, respectively. Alignment length and different statistics about the detected homologs are listed in the last columns. ORFans, datasets containing less than 20 homologs or with only one gene generating alternative isoforms are highlighted in light grey, blue and yellow, respectively.
(PDF)

S4 Table. Trees length. For each of the 154 used datasets, the tree length obtained with each method is listed. The shortest are highlighted in blue. For three datasets (corresponding to proteins Q6ZN06, P58317 and Q8N8J6), the selection with Guidance failed due to program crash. They are highlighted in light orange.
(PDF)

S5 Table. DL scores. For each of the 154 used datasets, the DL score computed by Notung with each isoform selection strategy is listed. The most consistent are highlighted in blue. For

three datasets (corresponding to proteins Q6ZN06, P58317 and Q8N8J6), the selection with Guidance failed due to program crash. They are highlighted in light orange.
(PDF)

S1 Fig. Reference phylogenetic tree of eukaryotic species used for the gene trees reconciliations. This tree was built according to Lecointre and Le Guyader book [21], the Ensembl reference species tree for the metazoan part [22] and a personal communication from CBA.
(PDF)

Acknowledgments

The authors would like to acknowledge Dominique Guyot for his advice and expertise on parallel computing. The computational works have been performed on the LBBE/PRABI cluster. This work was supported by grants from the CNRS, the Institut Français de Bioinformatique to GP, the France Génomique consortium for GP, and the Région Rhône-Alpes PhD thesis grant 13-012735-01, which was funding the Ph.D of HP. CBA is a member of the Institut Universitaire de France and is funded by the "Ancestrome" project through the Agence Nationale de la Recherche grant ANR-10-BINF-01-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: HP GP.

Formal analysis: HP AS GP.

Funding acquisition: GP.

Investigation: HP.

Methodology: HP GP.

Project administration: GP CBA.

Resources: GP.

Software: HP AS GP.

Supervision: GP CBA.

Validation: GP CBA.

Visualization: HP.

Writing – original draft: HP GP.

Writing – review & editing: HP GP.

References

1. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res 18:1381–1392. https://doi.org/10.1101/gr.053678.106 PMID: 18669480
2. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. https://doi.org/10.1038/nature07509 PMID: 18978772

3. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA 102:10557–10562. https://doi.org/10.1073/pnas.0409137102 PMID: 16000407
4. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334 PMID: 10742046
5. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210. https://doi.org/10.1186/1471-2148-10-210 PMID: 20626897
6. Hughes AL, Friedman R (2007) The effect of branch lengths on phylogeny: an empirical study using highly conserved orthologs from mammalian genomes. Mol Phylogenet Evol 45:81–88. https://doi.org/10.1016/j.ympev.2007.04.022 PMID: 17574446
7. Zou Z, Zhang J (2015) Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol 32:2085–2096. https://doi.org/10.1093/molbev/msv091 PMID: 25862140
8. Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA 104:7489–7494. https://doi.org/10.1073/pnas.0701705104 PMID: 17449636
9. Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA et al. (2012) Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol Biol Evol 29:1837–1849. https://doi.org/10.1093/molbev/mss025 PMID: 22319161
10. Villanueva-Cañas JL, Laurie S, Albà MM (2013) Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol 5:457–467. https://doi.org/10.1093/gbe/evt017 PMID: 23377868
11. Penn O, Privman E, Landan G, Graur D, Pupko T (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767. https://doi.org/10.1093/molbev/msq066 PMID: 20207713
12. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835
13. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690
14. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147
15. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808 PMID: 9254330
16. Thompson JD, Plewniak F, Poch O (1999) A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 27:2682–2690. https://doi.org/10.1093/nar/27.13.2682 PMID: 10373585
17. Altschul SF, Madden TL, Schäffer AA Zhang J, Zhang Z, Miller W et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402 https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694
18. Sela I, Ashkenazy H, Katoh K, Pupko T (2015) Guidance2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res 43:7–14. https://doi.org/10.1093/nar/gkv318
19. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224. https://doi.org/10.1093/molbev/msp259 PMID: 19854763
20. Nguyen LT, Schmidt H A, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300 PMID: 25371430
21. Lecointre G, Le Guyader H (2006) The Tree of Life: A Phylogenetic Classification. Harvard University Press, Harvard.
22. Yates A, Akanni W, Amode M R, Barrell D, Billis K, Carvalho-Silva D et al. (2016) Ensembl 2016. Nucleic Acids Res 44:D710–716. https://doi.org/10.1093/nar/gkv1157 PMID: 26687719

23. Stolzer M, Lai H, Xu M, Sathaye D Vernot B, Durand D (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28:i409–415. https://doi.org/10.1093/bioinformatics/bts386 PMID: 22962460

24. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307–1320. https://doi.org/10.1093/molbev/msn067 PMID: 18367465

25. Stover BC, Muller KF (2010) TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:7. https://doi.org/10.1186/1471-2105-11-7 PMID: 20051126

26. Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10:1073–1095. PMID: 8412650