Choosing representative proteins based on splicing structure similarity improves the accuracy of gene tree reconstruction

Esaie Kuitche Kamela[1,*], Marie Degen[1], Shengrui Wang[1], Aïda Ouangraoua[1,*]

[1] Department of Computer Sicence, University of Sherbrooke, Quebec, J1K 2R1, Canada

*Correspondence to be sent to: Department of Computer Sicence, University of Sherbrooke, 2500 Boul. de l'Université Sherbrooke, Quebec, J1K 2R1, Canada. E-mail: esaie.kuitche.kamela@usherbrooke.ca;aida.ouangraoua@usherbrooke.ca

Abstract. Constructing accurate gene trees is important, as gene trees play a key role in several biological studies, such as species tree reconstruction, gene functional analysis and gene family evolution studies. The accuracy of these studies is dependent on the accuracy of the input gene trees. Although several methods have been developed for improving the construction and the correction of gene trees by making use of the relationship with a species tree in addition to multiple sequence alignment, there is still a large room for improvement on the accuracy of gene trees and the computing time. In particular, accounting for alternative splicing that allows eukaryote genes to produce multiple transcripts/proteins per gene is a way to improve the quality of multiple sequence alignments used by gene tree reconstruction methods. Current methods for gene tree reconstruction usually make use of a set of transcripts composed of one representative transcript per gene, to generate multiple sequence alignments which are then used to estimate gene trees. Thus, the accuracy of the estimated gene tree depends on the choice of the representative transcripts. In this work, we present an alternative-splicing-aware method called Splicing Homology Transcript (SHT) method to estimate gene trees based on wisely selecting an accurate set of homologous transcripts to represent the genes of a gene family. We introduce a new similarity measure between transcripts for quantifying the level of homology between transcripts by combining a splicing structure-based similarity score with a sequence-based similarity score. We present a new method to cluster transcripts into a set of splicing homology groups based on the new similarity measure. The method is applied to reconstruct gene trees of the Ensembl database gene families, and a comparison with current EnsemblCompara gene trees is performed. The results show that the new approach improves gene tree accuracy thanks to the use of the new similarity measure between transcripts. An implementation of the method as well as the data used and generated in this work are available at https://github.com/UdeS-CoBIUS/SplicingHomologGeneTree/.

Keywords: Gene tree, Phylogenetics, Algorithms, Alternative splicing, Splicing homology, Fuzzy C-means

1 Introduction

In the last two decades, several phylogeny methods based on multiple alignment of homologous sequences have been developed and used for gene tree reconstruction (eg. PhyML (Guindon and Gascuel 2003), RAxML (Stamatakis 2006), PhyloBayes (Lartillot and Philippe 2004)). It has been shown that the gene trees reconstructed solely based on multiple sequence alignments often contain many errors and uncertainties, due to numerous reasons which include the quality of gene annotations, the quality of multiple sequence alignments, the quantity of substitutions in the alignments, and the accuracy of the phylogeny reconstruction methods (Hahn 2007; Rasmussen and Kellis 2007). To overcome these limitations, the development of phylogenomics methods which make use a known species tree in addition to multiple sequence alignments has been fruitful to improve the accuracy of reconstructed gene trees (eg. TreeBeST (Schreiber et al. 2014), NOTUNG (Chen, Durand, and Farach-Colton 2000), ALE (Szöllősi et al. 2013), GSR (Åkerborg et al. 2009), TERA (Scornavacca, Jacox, and Szöllősi 2015), SPIMAP (Rasmussen and Kellis 2011), GIGA (Thomas 2010)). Several methods for gene tree correction guided by species tree have also been developed (eg. ProfileNJ (Noutahi et al. 2016), TreeFix (Wu et al. 2013), TreeFix-DTL (Bansal et al. 2015), LabelGTC (El-Mabrouk and Ouangraoua 2017), RefineTree(Lafond et al. 2013), and (Górecki and Eulenstein 2012)). Another category of methods aims at jointly inferring gene trees and species trees (eg. PHYLDOG (Boussau et al. 2013), DLRS (Sjöstrand et al. 2012), PhyloNet (Y. Wang and Nakhleh 2018), BP&P (Rannala and Yang 2017)). The later provide accurate reconstructions by modeling and resolving the discordance between gene trees and species trees (Maddison 1997), but they are also limited by a scalability problem for larger data sets. Despite the improvement in the accuracy and resolution of gene trees reconstructed by all these approaches, there is still a large room for improvement on the accuracy of gene trees and the computing time.

In this work, we approach the question of improving the quality of multiple sequence alignments used in phylogenomics methods for gene tree reconstruction, by choosing wisely the set of proteins representing genes of a gene family. A key factor for the accuracy of multiple sequence alignment is that aligned sequences should be homologs. It is well known that alternative splicing (AS) is a ubiquitous process in eukaryote organisms by which multiple transcripts and proteins are produced from a single gene (Keren, Lev-Maor, and Ast 2010; Nilsen and Graveley 2010; Blencowe 2006; Blencowe 2017; Modrek and Lee 2002; Kuitche, Lafond, and Ouangraoua 2017; Iñiguez and Hernández 2017). Gene tree reconstruction methods account for this information, by selecting and making use of one reference transcript/protein sequence per gene to reconstruct the gene trees that populate gene tree databases (eg. EnsemblCompra (Vilella et al. 2009), PhylomeDB (Huerta-Cepas et al. 2014), Panther (Mi, Muruganujan, and Thomas 2012), TreeFam (Schreiber et al. 2014)). Among existing gene tree databases, the EnsemblCompara method is the most popular one (Vilella et al. 2009). Like all current gene tree reconstruction methods, the EnsemblCompara method chooses the longest transcript for each gene as its representative transcript to generate a

multiple sequence alignment. This choice ignores the transcripts of all the other genes of the gene family. The rationale of this choice is that the longest transcript is the more representative because its coverage of the gene is the largest (Kasukawa et al. 2004). However, there is no guarantee that the set of representative transcripts composed of the longest transcript of each gene constitutes the set of the most homologous transcripts between genes. The homology level of two transcripts from two homologous genes can be measured by the quantity of homologous nucleotides and exons shared by the transcripts. The longest transcripts of two homologous genes may be composed of completely different exons. Thus, choosing the longest transcript for each gene, independently of other genes, may lead to poorly homologous transcripts, and then erroneous sequence alignments and gene trees. Ideally, the set of transcripts chosen to estimate a gene tree must be equivalogs (Haft et al. 2001). Equivalogs are homologous transcripts that have conserved their functions since their last common ancestor. Equivalogs share the same functions, as well as conserved sequences and homologous exons, which results in similar splicing structures.

Classical methods for clustering homologous transcripts into equivalog families and automated functional identification of proteins rely on sequence and folding structure similarity. This approach ignores the splicing structure similarity between transcripts (Haft et al. 2001; Pandit et al. 2002; Zhang and Skolnick 2005; Balamurugan, Dekker, and Waldmann 2005; Balaji and Srinivasan 2007; Xu and Zhang 2010). However, it has been shown that the splicing structure of homologous sequences is often more conserved than the nucleotide sequences (Betts et al. 2001; Abril, Castelo, and Guigó 2005; Poulos et al. 2011). It has also been shown that the splicing structure contains information that can improve protein and transcript comparison (Abascal, Tress, and Valencia 2015; Abascal, Ezkurdia, et al. 2015).

In this paper, we introduce a new framework for the reconstruction of eukaryote gene trees. We consider all transcripts of all genes of a gene family, and we define a new similarity measure by combining splicing structure-based similarity and sequence-based similarity. Based on the all-against-all similarity matrix between transcripts, we propose a clustering method, based on an improved version of the fuzzy c-means (FCM) algorithm (Bezdek, Ehrlich, and Full 1984) and a latent representation of transcripts. This method allows to obtain splicing homology groups, that represent estimated equivalog families. Finally, we use a multiple sequence alignment of the selected representative transcripts to build the gene tree.

The paper is organized as follows: the next section describes the method devised to estimate gene trees by taking into account the splicing structure of transcripts. The Results and Discussion section presents the results of a comparative analysis between the gene trees estimated using the new method, and the EnsemblCompara gene trees which are currently the most largely used. The trees reconstructed with our method compare favorably with the trees from the EnsemblCompara database, in terms of the significance of a likelihood differ-

ence computed using an Approximately unbiased test (Shimodaira 2002), and in terms of the reconciliation with a species tree.

2 Materials and Methods

We first introduce some formal definitions that will be useful for the description of the method. Next, we present the definition of the new measure of similarity between transcripts, and then the new algorithm for clustering transcripts from a similarity matrix. In the last part of this section, we present how the gene trees are estimated from a set of representative transcripts. An overview of the method is depicted in Figure 1.

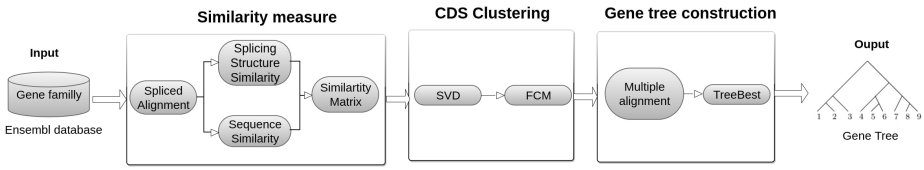

Fig. 1: **Overall view of the method**: The first step consists in the computation of the similarity matrix between Coding DNA Sequence (CDS) of transcripts. The second step consists in grouping CDS into splicing homology groups and, finally, the last step consists in aligning the representative CDS and building the gene tree.

2.1 Preliminary definitions

Gene and CDS: A gene sequence is a DNA sequence on the alphabet of nucleotides $\sigma = \{A, C, G, T\}$. Given a gene sequence g, an exon of g is a transcribed segment of g. A coding DNA sequence (CDS) of g is a concatenation of translated exons of g. The following is an example of toy gene g with two alternative CDS c_1 and c_2. The alignment of g , c_1 and c_2 shows that c_1 is composed of 3 exons, and c_2 of 4 exons.

```
>g
CGTATGGAATGCGTAAGAAGCAGGTCGTAAGCATACGTGGGTAAGGGGAATGATTGAAAG
>c1
ATGCAAGCAGGTCGGGAATGA
>c2
ATGGAATGCAAGCAGCATACGTGGGGGAATGATTGA

g : CGTATGGAATGCGTAAGAAGCAGGTCGTAAGCATACGTGGGTAAGGGGAATGATTGAAAG
c1: --------ATGC-----AAGCAGGTC-------------------GGGAATGA
c2: ---ATGGAATGC-----AAGCAG--------CATACGTGG-----GGGAATGATTGA
```

Multiple splicing structure: Given a set of CDS $\mathcal{C}$ from a set of homologous genes, a segment class of $\mathcal{C}$ is a set composed of exactly one, possibly empty, segment from each element of $\mathcal{C}$, such that these segments are homologous, i.e descending from the same ancestral segment. A segment class must contain at least one non-empty segment. A multiple splicing structure of a set of CDS $\mathcal{C}$ is a chain $E = \{E[1], ..., E[n]\}$ of segment classes of $\mathcal{C}$ in which no pair of classes $E[i]$ and $E[j]$ contains an overlapping segment, and each CDS c in $\mathcal{C}$ is the concatenation of its segments belonging to classes of E. Figure 2 presents an example of multiple spliced structure of three CDS decomposed into three segment classes. In practice, a set of homologous CDS is decomposed into a chain of segment classes corresponding to homologous exon classes.

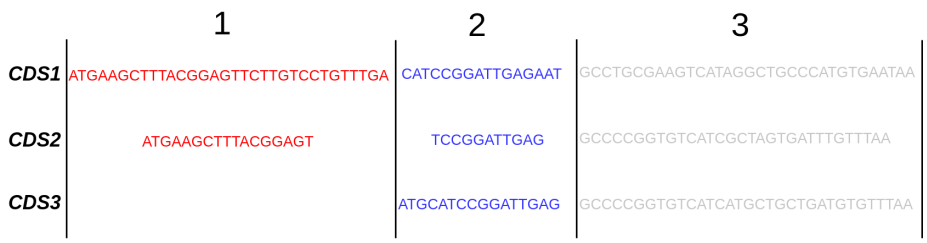

Fig. 2: **Multiple splicing structure**: Multiple splicing structure of three homologous CDS decomposed into three segment classes represented by different colors: red color for the first class, blue for the second class, and grey for the third class. Each CDS has a non-empty segment in all segment classes, except CDS3 which has an empty segment in the first class.

A multiple spliced alignment : Given a set of homologous CDS $\mathcal{C}$ decomposed into a multiple splicing structure $E = \{E[1], ..., E[n]\}$, the multiple spliced alignment of $\mathcal{C}$ corresponding to E is a chain $A = \{A[1], ..., A[n]\}$ such that each $A[i]$ is an alignment of the segment class $E[i]$. For any CDS c in $\mathcal{C}$, $A_c = \{A[i]_{|c} | 1 \leq i \leq n\}$ is the chain of n gapped segments of c induced by the multiple spliced alignment A. Given a multiple spliced alignment A, $|A|$ denotes the number of sub-alignments composing A, i.e. the number of segment classes in the corresponding multiple spliced alignment. Given a gapped nucleotide segment x, $len(x)$ denotes the number of nucleotides in x. Figure 3 gives an example of multiple spliced alignment corresponding to the multiple splicing structure of three CDS from Figure 2.

2.2 Similarity measure

The similarity measure between two CDS, x and y, of a set of homologous CDS is defined as follows:

$$Sim(x, y) = \alpha \times Struct(x, y) + (1 - \alpha) \times Seq(x, y)$$

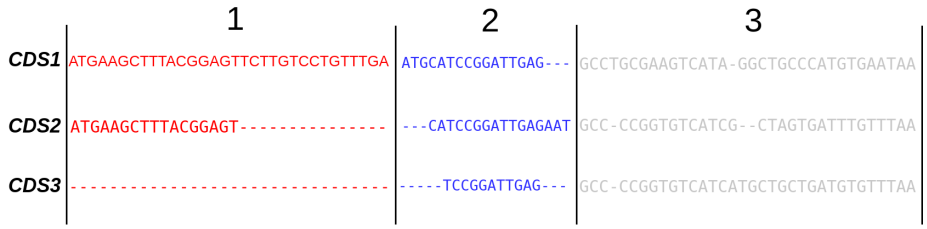

Fig. 3: **Multiple spliced alignment**: a multiple spliced alignment A corresponding to the multiple splicing structure of three CDS from Figure 2. For instance, $|A| = 3$, $len(A[1]_{|CDS3}) = 0$, and $len(A[2]_{|CDS3}) = 11$.

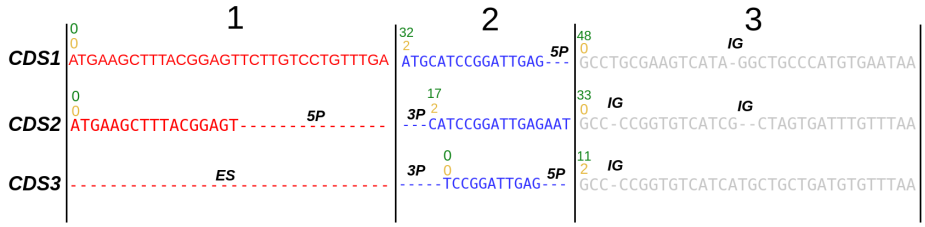

Fig. 4: **Alternative splicing events**: The multiple spliced alignment of Figure 3 with gapped segments annotated with their alternative splicing events, translation phases, and segment lengths. Above the first nucleotide of each segment, the number in green represents the position of the first nucleotide in the CDS and the number in orange represents the translation phase of the exon. The alternative splicing events of each segment are indicated (ES, 3P, 5P, or IG).

$Struct(x, y)$ is a splicing structure similarity between x and y, $Seq(x, y)$ is a sequence similarity between x and y, and α is such that $0 \leq \alpha \leq 1$ and defines the relative weights of $Struct(x, y)$ and $Seq(x, y)$ in the similarity score.

The value of $Struct(x, y)$ is defined by combining three similarity scores:

$$Struct(x, y) = \frac{SE(x, y) + TP(x, y) + SL(x, y)}{3}$$

$SE(x, y)$, $TP(x, y)$ and $SL(x, y)$ are respectively a splicing event (SE) similarity score, a translation phase (TP) similarity score, and a segment length (SL) similarity score between x and y, defined thereafter.

Splicing event similarity: Given a multiple spliced alignment A of a set of CDS $\mathcal{C}$, we consider four possible types of alternative splicing events for each gapped segment $A[i]_{|x}$ such that $1 \leq i \leq |A|$ and $x \in \mathcal{C}$. Figure 4 illustrates the different types of alternative splicing events. An exon skipping (ES) arises when the segment contains only gaps. For instance, the first segment of CDS3 in Figure 4 has an ES event. A 3' alternative splice site (3P) event arises when the segment starts with gaps. For instance, the second segments of CDS2 and CDS3 have a 3P event. A 5' alternative splice site (5P) event arises when the segment ends with gaps. For instance, the first segment of CDS2 as well as the second segments of CDS1 and CDS3 have a 5P event. An internal gap (IG) event arises when there is a set of consecutive gaps within the segment. For instance, the third segment of CDS1, CDS2 and CDS3 have at least one IG event.

For a CDS x, we denote by $e[i]_{|x}$ the set of alternative splicing events displayed by the gapped segment $A[i]_{|x}$. Given two CDS x and y from $\mathcal{C}$, the splicing event similarity between the two gapped segments $A[i]_{|x}$ and $A[i]_{|y}$ is the ratio of the number of common events in the two sets $e[i]_{|x}$ and $e[i]_{|y}$:
$SE(A[i]_{|x}, A[i]_{|y}) = \frac{|e[i]_{|x} \cap e[i]_{|y}|}{|e[i]_{|x} \cup e[i]_{|y}|}$
The splicing event similarity between x and y is then given by:
$SE(x, y) = \sum_{i=1}^{i=|A|} \frac{SE(A[i]_{|x}, A[i]_{|y})}{|A|}$

Translation phase similarity: Given a multiple spliced alignment A of a set of CDS $\mathcal{C}$, for each gapped segment $A[i]_{|x}$ such that $1 \leq i \leq |A|$ and $x \in \mathcal{C}$, we denote by $phase[i]_{|x}$ the translation phase of the segment $A[i]_{|x}$. The value of $phase[i]_{|x}$ equals the position of the first nucleotide of $A[i]_{|x}$ in the sequence x modulo 3. Given two CDS x and y from $\mathcal{C}$, the translation phase similarity between the two gapped segments $A[i]_{|x}$ and $A[i]_{|y}$ is 1 if the two segments have the same translation phase, and 0 otherwise:
$TP(A[i]_{|x}, A[i]_{|y}) = \begin{cases} 1 \text{ if } phase[i]_{|x} = phase[i]_{|y} \\ 0 \text{ otherwise.} \end{cases}$
The translation phase similarity between x and y is then given by:
$TP(x, y) = \sum_{i=1}^{i=|A|} \frac{TP(A[i]_{|x}, A[i]_{|y})}{|A|}$

Segment length similarity: Given two CDS x and y from $\mathcal{C}$, the segment length similarity between the two gapped segments $A[i]_{|x}$ and $A[i]_{|y}$ is given by:
$SL(A[i]_{|x}, A[i]_{|y}) = 1 - \frac{|len(A[i]_{|x}) - len(A[i]_{|y})|}{len(A[i]_{|x}) + len(A[i]_{|y})}$
The segment length similarity between x and y is then given by:
$SL(x, y) = \sum_{i=1}^{i=|A|} \frac{SL(A[i]_{|x}, A[i]_{|y})}{|A|}$

Sequence similarity: Giving a multiple spliced alignment A of a set of CDS $\mathcal{C}$, we use the substitution matrix *BLOSUM100* as substitution model to compute the sequence similarity score between each pair of CDS in $\mathcal{C}$.

Based on the definition of the similarity score $Sim(x, y)$ between two CDS x and y, an all-against-all similarity matrix is computed for the set of homologous CDS $\mathcal{C}$ of any gene family.

2.3 CDS Clustering

In order to determine the representative CDS of a set of homologous genes, CDS are first clustered into splicing homology groups. It has been demonstrated that, for clustering problems where clusters may overlap each other, fuzzy clustering methods, which can assign a data to more than one cluster, achieve higher performance than hard clustering methods, which assign each data to a single cluster (Sun, S. Wang, and Jiang 2004). The CDS clustering problem is better to be approached as a fuzzy clustering problem because a CDS in one gene may be similar in sequence and splicing structure to several alternative CDS of another gene, and splicing homology groups could overlap.

The Fuzzy C-means (FCM) algorithm and its derivatives are the most widely used fuzzy clustering algorithm (Bezdek, Ehrlich, and Full 1984). As input, they require, a representation of the data, generally a vectorial representation, which facilitates computing of the centroid of a cluster of data. They also require as input the number of clusters in the given data set. The clustering method starts with a transformation of the CDS representation into a vectorial representation, and the application of the FCM-Based Splitting Algorithm (FBSA) (Sun, S. Wang, and Jiang 2004) to cluster CDS.

Vectorial representation of CDS: In this section, we propose to compute a latent vectorial representation of CDS from a $N \times N$ similarity matrix S, containing the pairwise similarity scores between the N CDS of a set of homologous genes. Such a representation allows CDS to be represented in a new geometric space while preserving their mutual similarity and facilitating computing of centroids in clustering algorithms. Each CDS will be represented as an N-dimensional vector by applying a spectral decomposition on the similarity matrix S. The spectral decomposition of a symmetric $N \times N$ matrix S is the decomposition of S into $S = UDU^T$, such that U is an orthogonal $N \times N$ matrix and D is a $N \times N$ diagonal matrix. The columns of U correspond to the eigen vectors of S, and the diagonal entries of D correspond to the eigen values.

We define an $N \times N$ matrix $X = U \times \sqrt{D}$, such that each row x_i of X is the new vectorial representation of the ith CDS from the similarity matrix S. This transformation is valid because the similarity score S_{ij} between the ith and j^{th} CDS is now approximately equal to the inner vector product $< x_i, x_j >$ of the vectors representing the CDS. Thus, the relative similarity scores between CDS are preserved by the transformation, and the vectors $\{x_1, \ldots, x_N\}$ can be clustered using a fuzzy clustering algorithm.

Application of the FBSA for fuzzy clustering:

The FCM algorithm requires, as input parameter, the number of clusters in the data-set in order to be clustered (Bezdek, Ehrlich, and Full 1984). Given N data vectors $X = \{x_1, x_2, ..., x_N\}$ to cluster, and c the expected number of clusters, the FCM algorithm searches for a $N \times c$ non-negative matrix M called the fuzzy partition matrix. Each value m_{ki} such that $1 \leq k \leq N$ and $1 \leq i \leq c$ in the matrix M is the membership value of vector x_k to the ith cluster, and for any vector x_k, $\sum_{i=1}^{c} m_{ki} = 1$. Given a $N \times c$ membership matrix M, the centre of each cluster i, $1 \leq i \leq c$, is computed as the vector:

$$v_i = \frac{\sum_{k=1}^{N} m_{ki}^f x_k}{\sum_{k=1}^{N} m_{ki}^f} 1$$

The exponent $f > 1$ in the cluster centre formula is a parameter called a fuzzifier. Given a set of cluster centres $V = \{v_1, \ldots, v_c\}$, the values of the $N \times c$ membership matrix M are computed as follows:

$$m_{ki} = \begin{cases} (\sum_{j=1}^{c} (\frac{\|x_k - v_i\|}{\|x_k - v_j\|})^{\frac{2}{f-1}}) & if \, \|x_k - v_i\| > 0, \forall j \\ 1 & if \, \|x_k - v_i\| = 0 \\ 0 & if \exists j \neq i \, \|x_k - v_j\| = 0 \end{cases} \quad 2$$

The formula (1) and (2) are derived from minimization of the following objective function F_f.

Given N data vectors $X = \{x_1, x_2, ..., x_N\}$ and an expected number of clusters c, the FCM algorithm computes a membership matrix M and a set of cluster centres $V = \{v_1, \ldots, v_c\}$ that minimizes the following function

$$F_f(U, V) = \sum_{k=1}^{N} \sum_{i=1}^{c} m_{ki}^f \|x_k - v_i\|^2 \, 3$$

The FCM algorithm starts by randomly initializing the cluster centres $V = \{v_1, \ldots, v_c\}$. Then it iteratively recalculates the membership matrix M using equation (2) and a new set of cluster centres using equation (1) until a set of stable cluster centres is reached.

The main limitation of the FCM algorithm is that it requires the number of clusters as input. Several derivatives of the FCM algorithm have been developed for automatically determining the number of clusters. As input parameters, they take a lower bound c_{min} and an upper bound c_{max} for the number of clusters, compute an optimal clustering using FCM for each number of cluster c such that $c_{min} \leq c \leq c_{max}$, and choose the best value of c based on a cluster validity criterion.

The FBSA algorithm is a derivative of the FCM algorithm that allows to cluster a data set while automatically determining the number of clusters (Sun, S. Wang, and Jiang 2004). FBSA improves on other derivatives of the FCM algorithm to allows automated selections of the number of clusters. It reduces the randomness in the initialization of cluster centres at the beginning of each

clustering phase. For each clustering phase with a value of c such that $c_{min} < c \leq c_{max}$, the clustering is carried out starting with the previously obtained $c-1$ clusters and splitting the worst cluster in two clusters to obtain c clusters. The worst cluster is the one corresponding to the minimum value of a score function $S(i)$ associated with each cluster i as follows:

$$S(i) = \frac{\sum_{k=1}^{N} m_{ki}}{number_of_data_vectors_in_cluster_i}$$

A second improvement of FBSA in the selection of the best number of clusters is the use of a new validity index $V_d(c)$ based on a linear combination of cluster compactness and separation that is efficient even when clusters overlap each other. The validity index $V_d(c)$ is computed for each cluster number c to choose the number of clusters that maximizes $V_d(c)$. The detailed description of FBSA and the definition of the validity index are provided in (Sun, S. Wang, and Jiang 2004).

Given a $N \times N$ similarity matrix S on a set of N homologous CDS $\mathcal{C}$ of a gene family, we extend the FBSA method to cluster $\mathcal{C}$ into subsets of splicing homology groups, and select a set of representative CDS as follows:

1. Apply a spectral decomposition on the matrix S to obtain a vectorial representation $X = \{x_1, \ldots, x_N\}$ of the N CDS in $\mathcal{C}$;
2. Apply FBSA on $X = \{x_1, \ldots, x_N\}$ and choose the best clustering corresponding to a number c_{opt} of clusters maximizing $V_d(c_{opt})$;
3. Select the cluster i in the c_{opt}-clustering such that $S(i)$ is maximal.
4. For each gene g, select as representative CDS, the CDS k of g such that the membership M_{ki} of x_k to the ith cluster is maximum.

2.4 Gene tree construction

In order to carry out a comparative study on the effect of the choice of representative CDS for gene tree construction, the new method uses the same steps as the Ensembl method, except for the representative CDS selection step. After selecting the representative CDS, both methods compute a multiple sequence alignment of the protein sequences corresponding to the representative CDS using Mafft (Katoh and Standley 2013). Next, the CDS back-translated protein alignment and the species tree are used to estimate the gene trees with the TreeBest phylogenetic reconstruction method. TreeBest is a method for building gene trees using a known species tree (Schreiber et al. 2014).

3 Results and Discussion

In order to evaluate the new gene tree reconstruction method, we compare it with the method from EnsemblCompara, based on a dataset of gene families from the EnsemblCompara database (Vilella et al. 2009). In the remaining section,

since the differences between the two compared methods lies within the method used to select a set of representative transcripts for genes, the methods are named according to the latter. The EnsemblCompara method that selects the longest transcript of each gene is named the Longest Transcript (`LT`) method, and the new method that computes and uses a splicing homology group is called the Splicing Homology Transcript (`SHT`) method.

3.1 Dataset and compared methods

The initial dataset contains 225 gene families selected randomly from the EnsemblCompara database (release 96) (ibid.). The data for each gene family is composed of a set of homologous genes, their Coding Dna Sequences (CDS) and the gene tree from the EnsemblCompara database called the `LT_bd` gene tree. For each gene, only CDS that start with a codon start and whose lengths are multiple of 3 are considered, because the information on the coding phase is required for computing the similarity scores between CDS in the first step of the method, and the translations into protein sequences are needed for building the gene tree using TreeBest (Schreiber et al. 2014) in the last step of the method. Thus, the LT_bd gene tree contains only genes for which at least one CDS is considered. The LT_bd gene tree is induced from the EnsemblCompara database gene tree that is often on a larger set of genes.

Figure 5 shows some statistics on the number of genes per family and the number of CDS per gene in the dataset. Figure 5 (a) shows the percentage of gene families per ratio of the number of genes retained in this study over the number of genes in the EnsemblCompara database. For instance, we observe that in more than 50 % of gene families, the number of genes used in this study is less than 10 % of the number of genes in the EnsemblCompara gene family. Figure 5 (b) shows the percentage of genes per number of CDS. We observe that almost 80% of genes have only one CDS, and 15% of genes have two CDS. Figure 5 (c) shows the percentage of gene families per percentage of genes having a single CDS. We note that more than 55% of gene families have more than 90% of genes having a single CDS. Out of 225 gene families, 97 families have a single CDS per gene, which induces a single possible set of representative transcripts, thus the same tree is reconstructed using the two methods. These families were discarded, leaving a dataset with 128 gene families for the comparison. We call this dataset the 128-dataset, and the initial dataset is called the 255-dataset. Figure 5 (d) and (e) show the statistics on the number of CDS per gene in the 128-dataset. For instance, we observe that the number of single-CDS genes in the dataset is still high (more than 60 %), but the percentage of gene families that have more than 90% single-CDS genes is lower (less than 5% of gene families).

Despite the large number of single-CDS genes in the dataset, the comparative analyses between the LT and SHT reveal significant differences between the results of the two methods, as shown in the sequel.

The LT method was applied on the gene families to reconstruct a set of gene trees called the `LT` gene trees. Note that, for the same gene family, the LT gene tree may differ from the LT_bd gene tree, because the LT_bd is induced from

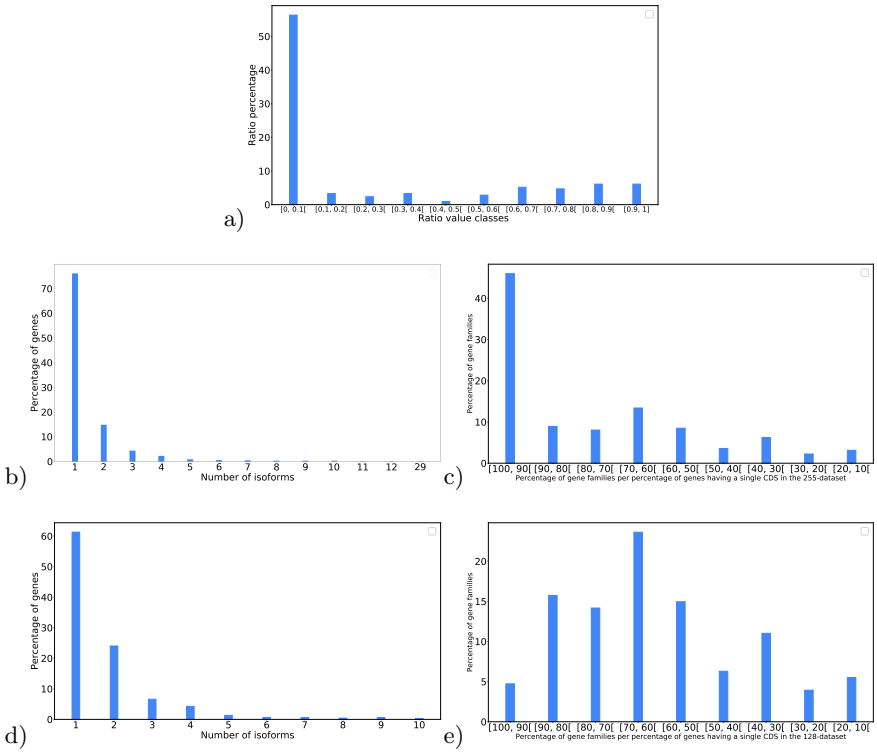

Fig. 5: **Statistics on the number of genes per family and the number of CDS per gene**: **(a)** Percentage of gene families per ratio of the number of genes retained in this study over the number of genes in the EnsemblCompara database. **(b)** Percentage of genes per numbers of CDS in the 255-dataset. **(c)** Percentage of gene families per percentage of genes having a single CDS in the 255-dataset. **(d)** Percentage of genes per numbers of CDS in the 128-dataset. **(e)** Percentage of gene families per percentage of genes having a single CDS in the 128-dataset.

the EnsemblCompara database gene tree that was often constructed on a larger set of genes as shown in Figure 5 (a). Using the SHT method, the gene trees were reconstructed for 7 different values of the parameter α of the SHT method, $\alpha = 0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0$, which corresponds to the weight of the splicing-structure similarity in the similarity scores between CDS. This yielded 7 sets of gene trees for the SHT method called the `SHT` gene trees. For each value of α, the resulting SHT gene trees were compared with the corresponding LT and LT_bd gene trees.

3.2 Comparison criteria

Three criteria are used to compare the SHT and LT gene trees for each value of $\alpha = 0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0$. The first criterion is the percentage of common transcripts between the sets of representative transcripts computed in the LT and SHT method. This criterion is used to evaluate the similarity between the methods in terms of the set of selected representative transcripts. The second criterion is the percentage of trees from each method which pass an approximately unbiased (AU) test (Shimodaira 2002) used to evaluate the significance of a likelihood difference between candidate trees on the same set of taxa. The last criterion is the reconciliation cost of the gene trees with the species trees.

3.3 Comparison based on the percentage of common representative transcripts

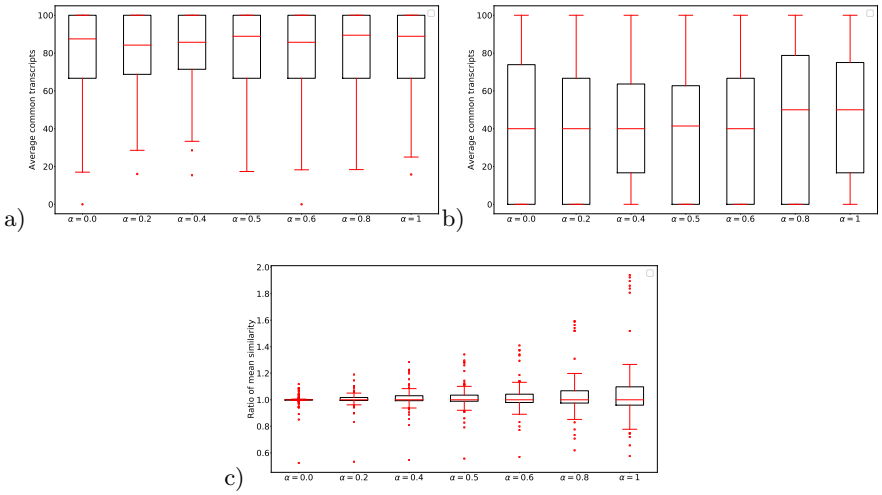

Fig. 6: **Common representative transcripts**: Boxplot of percentages of common representative transcripts in gene families between the SHT and the LT methods per values of α **(a)** in the 128-dataset, **(b)** when considering only genes having more than one transcript in the 128-dataset. **(c)** Boxplot of ratios of the mean similarity between the SHT transcripts over the mean similarity between the LT transcripts.

Figure 6 shows the percentage of common representative transcripts between the LT and SHT methods. We observe that the representative transcripts chosen

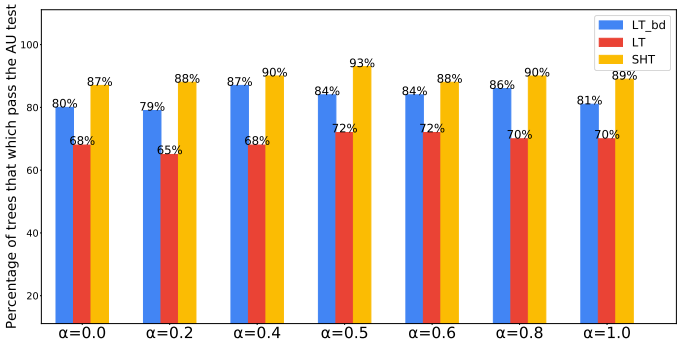

Fig. 7: **AU test**: Percentage of trees from each method (LT_bd, LT, SHT) which pass the AU test per values of α.

by the SHT method are not always the longest ones. Figure 6 (a) shows that for various values of α, the median percentage of common representative transcripts in gene families ranges between 80% and 84% for the 128-dataset. Figure 6 (b) shows that when we consider only genes having more than one transcript, the median percentage of common representative transcripts decreases and ranges between 42% and 48%. Figure 6 (c) shows the average ratio of the mean similarity between the SHT transcripts over the mean similarity between the LT transcripts. As expected, we observe that the SHT transcripts are closer in terms of similarity than the LT transcripts. These results illustrate that the SHT method is effective at computing a set of representative transcripts different from the longest transcripts selected by the LT. The SHT representative transcripts are closer in terms of similarity than the LT transcripts are when we increase the weight α of the splicing structure similarity score.

We also observe very few changes between the results of various values of α that determine the weights of the splicing structure similarity and the sequence similarity scores in the CDS similarity measure. It suggests that there is a correlation between the splicing structure similarity and the sequence similarity scores. This is confirmed by the mean of the correlation scores between the two scores computed for all gene families of the 255-dataset: minimum correlation score equals 0.21, maximum 1.0, mean 0.62, and median 0.60.

3.4 Comparison based on the percentage of trees passing the AU test

We used the approximately unbiased (AU) test to compare the trees built by the LT and the SHT methods (Shimodaira 2002). The AU test compares the likelihood of a set of trees in order to build a confidence set, which will consists of the trees that cannot be statistically rejected as a valid hypothesis. We say a tree can be rejected if its AU value, interpreted as a $p - value$, is under 0.05.

Otherwise, no significant evidence allows us to reject one of the two trees. We start by executing PhyML (release 27) (Guindon and Gascuel 2003) on the trees to obtain the log-likelihood values per site and we execute Consel (release 28) to obtain the results from the AU test.

Figure 7 shows the percentage of LT_bd, LT and SHT trees passing the AU test for various values of α. We clearly observed that the SHT method outperforms the LT_bd and the LT method. Between 88% to 93% of the SHT gene trees pass the AU test. 93% of trees passing the AU test are obtained for $\alpha = 0.5$. Whereas, for the LT_bd trees, the results vary between 79% and 87%, and for LT trees between 65% and 72%. Thus, the SHT method computes more trees which pass the AU test than the LT method. The SHT gene trees also allow to identify several cases in which the LT gene trees do not pass the AU test. It confirms that the choice of the longest transcript is not always the best approach to select a set of representative transcripts. It supports the use of the new similarity measure and clustering method to select sets of representative transcripts more wisely. The results also allow to compare the accuracy of the LT and LT_bd gene trees. They show that using a larger set of genes provides higher accuracy for the LT_bd gene trees as compared to the LT trees that are computed on a restricted set of genes.

3.5 Comparison based on the reconciliation cost with a species tree

Figure 8 (a) shows the comparative results between the SHT and the LT gene trees in terms of reconciliation cost on the 128-dataset. Figure 8 (b) shows the comparative results between the SHT and the LT_bd gene trees on the 225-dataset.

These results show that the LT_bd and SHT trees are identical in most of cases (between 70% and 73% of cases for various values of α), while the LT trees differs more often from the SHT trees. This result, coupled with the results of Figure 7 on the three methods' accuracy comparison, (LT $<$ LT_bd $<$ SHT) shows that choosing wisely the set of representative transcripts (i.e SHT trees) allows to reach the accuracy obtained using a larger set of genes (i.e LT_bd trees), and even to compute more accurate gene trees in several cases (Figure 7).

4 Conclusion

A new approach called Splicing Homology Transcript (SHT) method has been devised to estimate eukaryote gene trees. It is based on a novel and alternative way to choose the representative transcripts of a set of homologous genes. Contrary to current approaches that select the longest transcript of each gene as the representative transcript independently of other homologous genes, it selects a set of representative transcripts using a combination of the sequence similarity and the splicing structure similarity between sequences. This approach yields

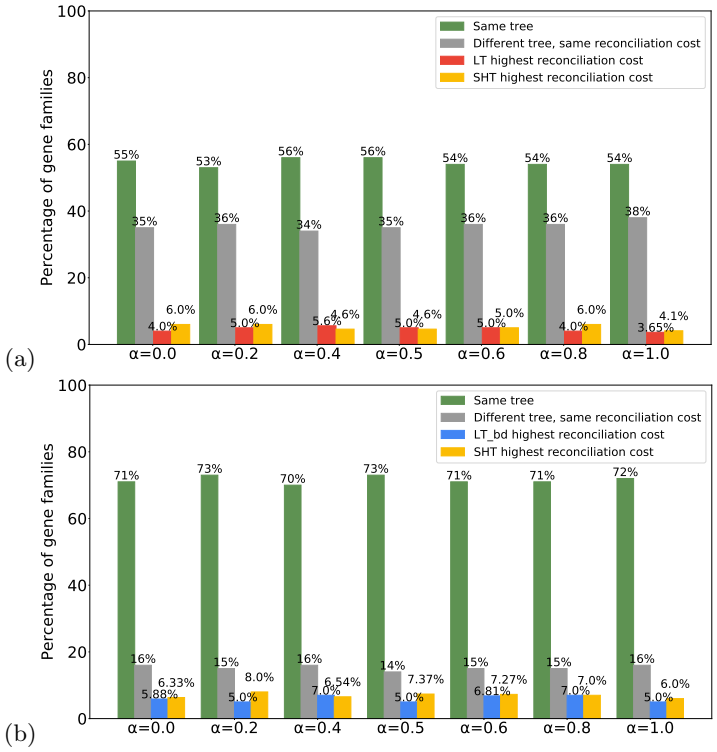

Fig. 8: **Reconciliation cost**: **(a)** Percentage of gene families for which the SHT and the LT gene trees are identical (green), for which the SHT and the LT gene trees differ but that they have the same reconciliation cost with the species tree (grey), for which the LT gene tree has a lower reconciliation cost (red), and for which the SHT gene tree has a lower reconciliation cost (yellow), per values of α. **(b)** Same for the comparison between the SHT and the LT_bd gene trees.

more accurate gene trees that the approach that selects the longest transcript of each gene as the representative transcript. We predict that alternative-splicing-aware gene tree reconstruction methods will be instrumental to advancing our understanding of gene family and genome evolution.

The analysis of the sequence and the structure similarity scores between transcripts of a gene family shows that the two scores are correlated. Future works will explore further definitions of the structure similarity score and the use of the structure similarity score only for choosing accurate sets of representative transcripts for gene tree reconstruction. Moreover, the SHT method differs from the EnsemblCompara gene tree estimation method only by the strategy used in the representative transcripts's selection step. Future works will consider alternative strategies in the other steps, such as the multiple sequence alignment and

the phylogeny estimation steps. The method also selects a single set of representative transcripts while there can be several co-optimal sets of representative transcripts. Future extensions of the method will include the computation of a transcript tree for each co-optimal set of representative transcripts, and the combination of the resulting trees to increase the precision of the final gene tree. Finally, a further extension of this work will consist of combining the set of representative transcript trees into a transcript phylogeny representing the evolutionary history of all transcripts of a set of homologous genes.

5 Acknowledgements

The authors thank the CoBiUS lab at University of Sherbrooke for their helpful constructive discussions.

References

Abascal, Federico, Iakes Ezkurdia, et al. (2015). "Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level". In: *PLoS computational biology* 11.6, e1004325.

Abascal, Federico, Michael L Tress, and Alfonso Valencia (2015). "The evolutionary fate of alternatively spliced homologous exons after gene duplication". In: *Genome biology and evolution* 7.6, pp. 1392–1403.

Abril, Josep F, Robert Castelo, and Roderic Guigó (2005). "Comparison of splice sites in mammals and chicken". In: *Genome research* 15.1, pp. 111–119.

Åkerborg, Örjan et al. (2009). "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis". In: *Proceedings of the National Academy of Sciences* 106.14, pp. 5714–5719.

Balaji, S and N Srinivasan (2007). "Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution". In: *Journal of biosciences* 32.1, pp. 83–96.

Balamurugan, Rengarajan, Frank J Dekker, and Herbert Waldmann (2005). "Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC)". In: *Molecular BioSystems* 1.1, pp. 36–45.

Bansal, Mukul S et al. (2015). "Improved gene tree error correction in the presence of horizontal gene transfer". In: *Bioinformatics* 31.8, pp. 1211–1218.

Betts, Matthew J et al. (2001). "Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution?" In: *The EMBO journal* 20.19, pp. 5354–5360.

Bezdek, James C, Robert Ehrlich, and William Full (1984). "FCM: The fuzzy c-means clustering algorithm". In: *Computers & Geosciences* 10.2-3, pp. 191–203.

Blencowe, Benjamin J (2006). "Alternative splicing: new insights from global analyses". In: *Cell* 126.1, pp. 37–47.

— (2017). "The relationship between alternative splicing and proteomic complexity". In: *Trends in biochemical sciences* 42.6, pp. 407–408.

Boussau, Bastien et al. (2013). "Genome-scale coestimation of species and gene trees". In: *Genome research* 23.2, pp. 323–330.

Chen, Kevin, Dannie Durand, and Martin Farach-Colton (2000). "Notung: dating gene duplications using gene family trees". In: *Proceedings of the fourth annual international conference on Computational molecular biology*, pp. 96–106.

Górecki, Pawel and Oliver Eulenstein (2012). "Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem". In: *BMC bioinformatics*. Vol. 13. S10. Springer, S14.

Guindon, Stéphane and Olivier Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood". In: *Systematic biology* 52.5, pp. 696–704.

Haft, Daniel H et al. (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins". In: *Nucleic acids research* 29.1, pp. 41–43.

Hahn, Matthew W (2007). "Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution". In: *Genome biology* 8.7, R141.

Huerta-Cepas, Jaime et al. (2014). "PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome". In: *Nucleic acids research* 42.D1, pp. D897–D902.

Iñiguez, Luis P and Georgina Hernández (2017). "The evolutionary relationship between alternative splicing and gene duplication". In: *Frontiers in genetics* 8, p. 14.

Kasukawa, Takeya et al. (2004). "Construction of representative transcript and protein sets of human, mouse, and rat as a platform for their transcriptome and proteome analysis". In: *Genomics* 84.6, pp. 913–921.

Katoh, Kazutaka and Daron M Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". In: *Molecular biology and evolution* 30.4, pp. 772–780.

Keren, Hadas, Galit Lev-Maor, and Gil Ast (2010). "Alternative splicing and evolution: diversification, exon definition and function". In: *Nature Reviews Genetics* 11.5, p. 345.

Kuitche, Esaie, Manuel Lafond, and Aıda Ouangraoua (2017). "Reconstructing protein and gene phylogenies using reconciliation and soft-clustering". In: *Journal of bioinformatics and computational biology* 15.06, p. 1740007.

Lafond, Manuel et al. (2013). "Gene tree correction guided by orthology". In: *BMC bioinformatics*. Vol. 14. S15. Springer, S5.

Lartillot, Nicolas and Hervé Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process". In: *Molecular biology and evolution* 21.6, pp. 1095–1109.

El-Mabrouk, Nadia and Aıda Ouangraoua (2017). "A general framework for gene tree correction based on duplication-loss reconciliation". In: *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Maddison, Wayne P (1997). "Gene trees in species trees". In: *Systematic biology* 46.3, pp. 523–536.

Mi, Huaiyu, Anushya Muruganujan, and Paul D Thomas (2012). "PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees". In: *Nucleic acids research* 41.D1, pp. D377–D386.

Modrek, Barmak and Christopher Lee (2002). "A genomic view of alternative splicing". In: *Nature genetics* 30.1, p. 13.

Nilsen, Timothy W and Brenton R Graveley (2010). "Expansion of the eukaryotic proteome by alternative splicing". In: *Nature* 463.7280, p. 457.

Noutahi, Emmanuel et al. (2016). "Efficient gene tree correction guided by genome evolution". In: *PLoS One* 11.8.

Pandit, Shashi B et al. (2002). "SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes". In: *Nucleic acids research* 30.1, pp. 289–293.

Poulos, Michael G et al. (2011). "Developments in RNA splicing and disease". In: *Cold Spring Harbor perspectives in biology* 3.1, a000778.

Rannala, Bruce and Ziheng Yang (2017). "Efficient Bayesian species tree inference under the multispecies coalescent". In: *Systematic biology* 66.5, pp. 823–842.

Rasmussen, Matthew D and Manolis Kellis (2007). "Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes". In: *Genome research* 17.12, pp. 1932–1942.

— (2011). "A Bayesian approach for fast and accurate gene tree reconstruction". In: *Molecular Biology and Evolution* 28.1, pp. 273–290.

Schreiber, Fabian et al. (2014). "TreeFam v9: a new website, more species and orthology-on-the-fly". In: *Nucleic acids research* 42.D1, pp. D922–D925.

Scornavacca, Celine, Edwin Jacox, and Gergely J Szöllősi (2015). "Joint amalgamation of most parsimonious reconciled gene trees". In: *Bioinformatics* 31.6, pp. 841–848.

Shimodaira, Hidetoshi (2002). "An approximately unbiased test of phylogenetic tree selection". In: *Systematic biology* 51.3, pp. 492–508.

Sjöstrand, Joel et al. (2012). "DLRS: gene tree evolution in light of a species tree". In: *Bioinformatics* 28.22, pp. 2994–2995.

Stamatakis, Alexandros (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models". In: *Bioinformatics* 22.21, pp. 2688–2690.

Sun, Haojun, Shengrui Wang, and Qingshan Jiang (2004). "FCM-based model selection algorithms for determining the number of clusters". In: *Pattern recognition* 37.10, pp. 2027–2037.

Szöllősi, Gergely J et al. (2013). "Efficient exploration of the space of reconciled gene trees". In: *Systematic biology* 62.6, pp. 901–912.

Thomas, Paul D (2010). "GIGA: a simple, efficient algorithm for gene tree inference in the genomic age". In: *BMC bioinformatics* 11.1, p. 312.

Vilella, Albert J et al. (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates". In: *Genome research* 19.2, pp. 327–335.

Wang, Yaxuan and Luay Nakhleh (2018). "Towards an accurate and efficient heuristic for species/gene tree co-estimation". In: *Bioinformatics* 34.17, pp. i697–i705.

Wu, Yi-Chieh et al. (2013). "TreeFix: statistically informed gene tree error correction using species trees". In: *Systematic Biology* 62.1, pp. 110–120.

Xu, Jinrui and Yang Zhang (2010). "How significant is a protein structure similarity with TM-score= 0.5?" In: *Bioinformatics* 26.7, pp. 889–895.

Zhang, Yang and Jeffrey Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic acids research* 33.7, pp. 2302–2309.