Current strategic limitations of phylogenetic tools badly impact the inference of an evolutionary tree

Shamantha Nasika, Ashish Runthala*

Email: ashish.runthala@gmail.com

Department of Biotechnology, Koneru Lakshmaiah Education Foundation, Vaddeswaram

Abstract: For drawing an evolutionary relationship among several protein sequences, the phylogenetic tree is usually constructed through maximum likelihood-based algorithms. To improve the accuracy of these methodologies, many parameters like bootstrap methods, correlation coefficient and residue-substitution models are presumably over-ranked to derive biologically credible relationships. Although the accuracy of protein sequence alignment and the substitution matrix are preliminary constraints to define the biological accuracy of the overlapped sequences/residues, the alignment is not iteratively optimized through the statistical testing of residue-substitution models. The study majorly highlights the potential pitfalls that significantly affect the accuracy of an evolutionary protocol. It emphasizes the need for a more accurate scrutiny of the entire phylogenetic methodology. The need of iterative optimizations is illustrated to construct a biologically credible and not mathematically optimal tree for a sequence dataset.

Keywords: Phylogenetic tree, substitution model, IQ-Tree, bootstrap, log-likelihood, IToL

Introduction

An evolutionary relationship is usually drawn to screen the sequence/functional similarity among the protein sequences (Chan 1989; Kamjula et al. 2020; Keeling et al. 1998; Sharma et

al. 2018; T. Lanisˇnik Rizˇner 2001). On the basis of mutual sequence similarity, the phylogenetic tree is constructed as the schematic depiction of the evolutionary connection between the species of different origins, and evolutionary distance/relatedness is assessed through several non-absolute scores including branch length, maximum likelihood score, and topology.

To draw the evolutionary relationships, a sequence profile or multiple sequence alignment (*MSA*) is usually constructed through several protocols, viz. MAFFT, MUSCLE, Kalign, TCOffee, Clustal-Omega and PROMALS3D (Di Tommaso et al. 2011; Edgar 2004; Kazutaka Katoh 2002; Lassmann and Sonnhammer 2005; Pei and Grishin 2014; Sievers et al. 2011). Multiple sequence comparison by log-expectation (MUSCLE; <https://www.ebi.ac.uk/Tools/msa/muscle/>) and multiple alignments using fast fourier transform (MAFFT; <https://www.ebi.ac.uk/Tools/msa/mafft/>) protocols are better than ClustalW and quickly align hundreds of sequences within seconds (Edgar 2004; Kazutaka Katoh 2002). Kalign alignment is also a very fast MSA tool which is applicable for large alignments. It uses WU-Manber string-matching algorithm to quickly align the sequences (Kazutaka Katoh 2002; Lassmann and Sonnhammer 2005). However, the profile multiple alignment with predicted local structures and topological constraints (PROMALS3D) proficiently aligns sequences/structures on the basis of sequence profiles, predicted secondary structures and structural constraints (<http://prodata.swmed.edu/promals3d/promals3d.php>). In contrast to these servers, Clustal-Omega is capable of aligning numerous sequences on the basis of a guide tree (Sievers et al. 2011) and hidden markov model (*HMM*) profiles (Sievers and Higgins 2018). Residue substitution matrix is used to align the sequences, and is deployed by the phylogeny protocol for deriving the evolutionary relationships. Phylogeny is routinely derived through

nine different substitution matrices, viz. Dayhoff (Winona C. Barker 1978), JTT, BLOSUM62, WAG, PMB, DCMut (Kosiol and Goldman 2005), JTTDCmut (Kosiol and Goldman 2005), LG (Le and Gascuel 2008), and VT (Alva et al. 2016).

On the basis of sequence alignment, evolutionary relationship is usually drawn through several tools including a tree and reticulogram reconstruction (*T-REX*), MEGA5, PhyML, randomized accelerated maximum likelihood (RAxML), NGPhylogeny.fr and IQTree. PhyML uses the maximum likelihood method along with bootstrap and other parameters (Guindon et al. 2005). While T-REX deploys maximum likelihood/parsimony methods (Boc et al. 2012), MEGA5 and NGPhylogeny.fr offer a customizable platform (Hall 2013; Lemoine et al. 2019). However, the RAxML protocol is routinely used for large sequence datasets (Stamatakis 2014). In comparison to these strategies, IQTree (Nguyen et al. 2015) offers a highly customizable protocol to build even the complex trees more accurately. The method uses the ultrafast bootstrap algorithm and saves the runtime through statistical scoring, and is significantly faster than RAxML and PhyML (Hoang et al. 2018; Minh et al. 2013).

Evolutionary analysis has been extensively used for various research projects since 1883, and until December 18, 2019, 207863 articles are published, ~1540 articles every year (BROOKS 1883; Xuan et al. 2019). Almost all phylogenetic algorithms are maximally dependent on the aforementioned parameters and the impact of a biologically inaccurate sequence alignment and an incompatible residue substitution matrix is greatly disregarded. Most of these methods deploy a default substitution/alignment protocol for every sequence dataset (Boc et al. 2012; Guindon et al. 2005; Hall 2013; Lemoine et al. 2019; Nguyen et al. 2015; Stamatakis 2014), and do not train the dataset over available options. Moreover, no previous study has simultaneously assessed the quantified the negative impact of an incorrect substitution matrix,

alignment protocol and the depth of a sequence dataset over the accuracy of a phylogenetic tree. By providing a comprehensive assessment of all the alignments and substitution models over an evolutionary tree, the study proves the importance of well-trained substitution model and biologically correct alignment. To aid the construction of a more accurate evolutionary server, the study defines the vital constraints for a heuristic and iterative tailoring of input data. It will aid us extract the closest template structures for consistently predicting the near-native protein models over the conventional protocols (Ashish and Shibasish 2013; Runthala 2012; Runthala 2015; Runthala 2020; Runthala and Chowdhury 2014; Runthala and Chowdhury 2016; Runthala and Chowdhury 2019), and will allow us extract a more-accurate functional relationship among the candidate sequences for various research domains (Amrein et al. 2019; Anil Kumar Jamithireddy 2020; Ashish Runthala 2011; Phulara et al. 2020).

Materials and methods

Building the sequence dataset and evaluating the sequence identity

For rigorously doing the evolutionary analysis through several parameters and to save the computational time, the families with less than 1000 sequences are prioritized. As the accuracy of a phylogeny protocol is strongly dependent on the number of sequences and the sequence length, three non-redundant datasets 1-3, containing 36, 109 and 712 sequences, are orderly derived from three randomly-selected families PF18258, PF18760 and PF00074 (Figure 1). As simulating the sequence datasets on basis of various constraints might lead to a biased perturbation of their individual and mutual diversity, the three arbitrarily defined datasets should be robustly accurate and more fruitful for the study. As sequence identity is an important basis of evolutionarily clustering the sequences for estimating their sequence divergence, the

mutual identity matrix is computed for each of these datasets through the Clustal-omega protocol (Sievers et al. 2011).

Figure 1: Flowchart representation of the methodology deploying the construction of three varied sequence datasets and the evolutionary analysis of their several alignments through diverse substitution models.

Construction of phylogenetic trees

The constructed sequence datasets are aligned through Clustal-Omega, Kalign, MAFFT, MUSCLE, TCOFFEE and Promals3D, and the HMM-based Clustal-Omega (*Clustal1*), Kalign (*Kalign1*), MAFFT (*MAFFT1*), MUSCLE (*MUSCLE1*) and TCOFFEE (*TCOFFEE1*) protocols

(Alva et al. 2016) to screen the more accurate alignment on the basis of phylogenetic tree. For the computational limit of 2500-character length of an input alignment, TCoffee protocol for both EBI and HHPred servers could not be deployed to align the set2 and set3 datasets. Likewise, for the sequence-strength limit of 500 entries, the MUSCLE alignment of EBI server is excluded for set3. A set of nine substitution models viz. Dayhoff (Winona C. Barker 1978), JTT, BLOSUM62, WAG, PMB, DCMut (Kosiol et al. 2006; Kosiol and Goldman 2005), JTTDCmut (Kosiol et al. 2006; Kosiol and Goldman 2005), LG(Le and Gascuel 2008), and VT (Alva et al. 2016) are likewise screened to decipher the more compatible matrix for a sequence dataset.

To quickly construct the evolutionary tree through an ultrafast bootstrap methodology on the basis of a customized protocol, the IQ-Tree server is used (Nguyen et al. 2015). The maximum likelihood based phenetic analysis parametrically estimates the evolutionary tree through several features including branch length, nucleotide composition bias and corrected distances between taxa (Nabhan and Sarkar 2012). As the theoretical probability is dependent on the observed data and not the experimental results, the correlation coefficient is further evaluated to find more accurate phylogenetic solutions (Bruce Rannala 1996; Kensche et al. 2008; Nunes et al. 2015). For each of the constructed sequence alignments, the evolutionary trees are derived through IQTree at the correlation coefficient of 0.9. As bootstrapping assesses the robustness of a phylogenetic solution (Felsenstein 1985), a quick 1000-bootstrap run is deployed to decipher the biologically-feasible substitution matrix and alignment protocol for each of the sequence datasets.

Evolutionary analysis

The maximum-likelihood based tree is assessed through log-likelihood, total tree length and the total internal branch length score. Evaluating the evolutionary trees, constructed through every alignment on basis of all the deployed substitution models, the most-accurate tree is defined to be the one with the lowest log-likelihood score. The mutually synchronous behavior of the other two scores is subsequently evaluated for a considered set of alignment and substitution model, as an attempt to showcase the current pitfalls of these measures and to screen a more accurate assessment protocol for all the datasets. The evolutionary trees are lastly visualized through ITOL (<https://itol.embl.de/>) for rationalizing the differences across various subfamilies (Sagulenko et al. 2018).

Results and Discussion

Building the sequence dataset

Despite purging the redundant hits, the datasets 1-3 orderly encompass 36, 109 and 712 sequences within the length range of 96-484, 84-4427 and 50-570 and encode a total of 6807, 150301 and 827623 residues respectively. It substantially increases the sequence space of these sets and only a few algorithms are found capable of building their alignments.

Unfortunately, no substitution matrix consistently yields the most accurate and biologically reasonable alignments for all sequence datasets (Kemena and Notredame 2009), and as an attempt to draw a more accurate evolutionary relationship, most of the phylogeny protocols use the bootstrap methodology (Huson and Bryant 2006). Although presently taken to be key score for finding the best evolutionary model for a sequence dataset, the log-likelihood score may not always select the most-accurate tree always and it should be strengthened with added measures

to make the analysis more accurate. An evolutionarily closer sequence, often sharing a higher sequence identity, should have a smaller branch length, and this should strengthen the approximation of the log-likelihood metric (Parks and Goldman 2014). While the datasets 1, 2 and 3 show a sequence identity within the range of 0-99.52, 0-97.97 and 0-99.53 respectively, the average sequence identity is orderly found to be of 27.452 ± 21.586 , 16.742 ± 7.815 and 29.137 ± 12.759 , and it suggests a higher evolutionary divergence among the set2 sequences.

Evolutionary analysis for three datasets

The maximum-likelihood based phylogenetic tree is constructed through the ultrafast bootstrap methodology of IQTree to save computational resources and time (Nguyen et al. 2015). The number of free parameters (*branch + model parameters*) for the three sets in all the alignments is orderly found to be 88, 234, 1440. For the defined sequence sets and all of their constructed trees, the assessment is done through the log-likelihood score (Chatzou et al. 2016; Parks and Goldman 2014), total tree length (Nguyen et al. 2015), and the sum of the internal branch lengths (Chang et al. 2014).

Correlating the maximum, minimum, average and standard deviation scores for the 13 constructed Clustal-OMEGA, MUSCLE, KALIGN, MAFFT, TCOFFEE1, ESPRESSO TCOFFEE, MCOFFEE, PSI TCOFFEE, MUSCLE1, Clustal1, MAFFT1, KALIGN1 and PROMALS3D alignments against the 9 substitution models, viz. BLOSUM62, Dayhoff, JTT, JTT DCMut, LG, PMB, VT, WAG, DCMut for set1, it is observed that the total tree length and total internal branch length follow a similar scoring pattern, as expected (Figure 2). However, the scoring undulation for the log-likelihood score is found quite different, and it falls in line with the recent publication (Sievers and Higgins 2018). A similar trend is observed for the set2

and set3 datasets. The residue substitution scores are plotted for the constructed alignments (Figure 3), wherein the X and Y axis orderly refers to substitution matrices and alignment protocols. For sets1-3, the total internal branch length and total tree length are found lowest for BLOSUM62 for the Clustal alignment in contrast to the highest respective scores shown by DCMut matrix for the PROMALS3D alignment. It indicates that the structure-based sequence alignment should not be directly deployed for a study, unless explicitly evaluated for the selected dataset. Hence, to derive a more accurate evolutionary relationship, the biologically correct sequence alignment and residue substitution model should be used (Pearson 2013a).

Figure 2: Maximum, minimum, average and standard deviation scoring undulations for (A) Log-likelihood score (B) total tree length (C) total internal branch length for the deployed substitution models (blue) and alignments (red) for the 36-sequence set1.

Figure 3: Scoring alterations of the total tree length and internal branch length of the three sequence datasets set1-set3 for the phylogenetic trees constructed through the alignments on the basis of the selected substitution models.

Finding the top-ranked alignments and substitution models

For a protein sequence dataset, the evolutionary trees are drawn on the basis of 9 substitution matrices and 13 sequence alignments for sets 1-3 respectively, and assessed through the log-likelihood, total tree-length and total internal branch-length scores. The evolutionary likelihood of a phylogenetic tree is majorly assessed through three scoring measures viz. log-likelihood (Nikoh et al. 1994), total tree length and total internal tree length (Nikoh et al. 1994), and a solution with the lowest score is expected to be an optimal one. Among the 279 evolutionary trees, the top-ranked phylogenetic solutions, yielded by the best alignments (Table I) and the more accurate substitution models (Table II), are evaluated for the lower and upper limits of the three scores. The parameters are likewise evaluated for the overall average scores for each

alignment protocol and substitution model, and the consistently correct alignment (Table III) and the substitution model (Table IV) are screened.

		Log-likelihood Alignment protocol (Substitution model)	Total tree length Alignment protocol (Substitution model)	Total internal tree length Alignment protocol (Substitution model)
Set1	Lowest	-11898.259 MUSCLE(LG)	9.453 Kalign1 (PMB)	5.289 Kalign1 (PMB)
	Highest	-9982.515 Kalign1 (WAG)	20.227 MUSCLE (LG)	12.423 PROMALS3D (Dayhoff)
Set2	Lowest	-391294.242 MUSCLE (Dayhoff)	48.06 Kalign1 (PMB)	11.301 Kalign1(PMB)
	Highest	-281093.705 Kalign1 (LG)	272.449 MUSCLE (DCMut)	62.593 PROMALS3D (LG)
Set3	Lowest	-128789.188 PROMALS3D (LG)	129.367 Kalign1 (VT)	51.48 Kalign1 (PMB)
	Highest	-120287.5159 MUSCLE1 (JTT)	202.405 PROMALS3D (Dayhoff)	84.51 PROMALS3D (DCMut)

Table1: Lower and upper limit of the log-likelihood, total tree length and total internal tree length scores for the evolutionary tree of the top-ranked alignments, screened from the pool built through various substitution models of the three datasets.

A lower score indicates a better solution.

		Log-likelihood Best substitution model	Total tree length Best substitution model	Total internal tree length Best substitution model
Set1	Lowest	-11898.259 MUSCLE(LG)	9.453 Kalign1 (PMB)	5.289 Kalign1 (PMB)
	Highest	-9982.515 Kalign1 (WAG)	20.227 MUSCLE (LG)	12.423 PROMALS3D (Dayhoff)
Set2	Lowest	-391294.242 MUSCLE (Dayhoff)	48.06 Kalign1 (PMB)	11.301 Kalign1(PMB)

	Highest	-281093.705 Kalign1 (LG)	272.449 MUSCLE (DCMut)	62.593 PROMALS3D (LG)
Set3	Lowest	-128789.188 PROMALS3D (LG)	129.367 Kalign1 (VT)	51.48 Kalign1 (PMB)
	Highest	-120287.5159 MUSCLE1 (JTT)	202.405 PROMALS3D (Dayhoff)	84.51 PROMALS3D (DCMut)

Table2: Lower and upper limit of the log-likelihood, total tree length and total internal tree

length scores for the evolutionary tree constructed through the top-ranked substitution models

for the three datasets. A lower score indicates a better solution.

		Log-likelihood Alignment protocol	Total tree length Alignment protocol	Total internal tree length Alignment protocol
Set1	Lowest	-11800.865±51.083 MUSCLE	10.066±0.435 Kalign1	5.712±0.303 Kalign1
	Highest	-8986.079±7685.013 PSI Tcoffee	18.366±1.539 MUSCLE	10.978±1.511 PROMALS3D
Set2	Lowest	-390685.369±628.278 MUSCLE	51.254±2.24 Kalign1	12.113±0.515 Kalign1
	Highest	-285276.218±2031.102 Kalign1	229.908±38.641 MUSCLE	52.41±7.867 PROMALS3D
Set3	Lowest	-127100.2±914.485 MUSCLE1	179.305±7.585 MUSCLE1	77.491±3.421 MUSCLE1
	Highest	-121429.103±1105.85 PROMALS3D	190.593±9.087 PROMALS3D	80.327±3.580 PROMALS3D

Table3: Lowest and highest average score of the log-likelihood, total tree length and total

internal tree length scores for the evolutionary tree top-ranked alignments of the three

datasets. A lower score indicates a better solution.

		Log-likelihood Best substitution model	Total tree length Best substitution model	Total internal tree length Best substitution model
Set1	Lowest	-11376.241±487.047 LG	13.072±1.637 PMB	7.861±0.973 BLOSUM62
	Highest	-9474.36±6320.371 WAG	15.418±2.271 Dayhoff	9.513±1.54 Dayhoff
Set2	Lowest	-358941.659±29511.738 Dayhoff	112.637±39.248 PMB	30.21±9.284 PMB
	Highest	-357079.167±31845.036 WAG	155.0811±67.738 LG	41.579±15.52 Dayhoff
Set3	Lowest	-120369.906±4838.96 LG	159.124±14.712 PMB	65.974±7.549 PMB
	Highest	-117866.669±5195.16 JTT	175.996±18.798 Dayhoff	70.992±8.940 JTTDCMut

Table4: Lowest and highest average score of the log-likelihood, total tree length and total

internal tree length scores for the evolutionary tree constructed through the top-ranked

substitution models for the three datasets. A lower score indicates a better solution.

	Parameters	Best substitution matrix and alignment (Frequency)	Frequency (%)
Set1	Log-likelihood	LG-MUSCLE (12/13)	92.307
	Total tree length	PMB-Kalign1 (10/13)	76.923
	Total internal branch length	PMB-Kalign1 (10/13)	76.923
Set2	Log-likelihood	Dayhoff-MUSCLE (5/9)	55.556
	Total tree length	PMB-Kalign1 (7/9)	77.778
	Total internal branch length	PMB-Kalign1 (7/9)	77.778
Set3	Log-likelihood	LG-MUSCLE1 (4/8)	50
	Total tree length	PMB-PROMALS3D (4/8)	50
	Total internal branch length	PMB-MUSCLE1 (4/8)	50

Table 5: Top-ranked alignment-substitution model protocols for the three datasets

for the three assessment measures: Log-likelihood, total tree length and total

internal branch length. The failure to yield the consistently correct evolutionary

model for any sequence dataset is indicated.

Ranking the evolutionary trees to find the top-ranked alignment protocols for the three sets, it is observed that the MUSCLE, MUSCLE and PROMALS3D show the lowest log-likelihood score of -11898.259, -391294.242 and -128789.18 against the respective scores of -9982.515, -281093.705 and -120287.51 for the Kalign1, Kalign1 and MUSCLE1 alignments. Assessing with the minimal tree length and internal tree length measure, Kalign1, Kalign1 and PROMALS3D and Kalign1, Kalign1 and MUSCLE1 alignments are orderly found to construct better trees for sets1-3, against the respective solutions of MUSCLE, MUSCLE and MUSCLE1, and PROMALS3D, PROMALS3D and PROMALS3D (Edgar 2004; Pei et al. 2008). For the 9 low-scoring evolutionary trees for sets 1-3 for all the scoring measures, Kalign1, MUSCLE, MUSCLE1 and PROMALS3D are orderly found to yield the best solutions for 4, 2, 2 and 1 cases, and PMB, LG, JTT, Dayhoff models are found to yield the best solution for 5,1, 1, 1 and 1 cases (Arenas 2015; Pei and Grishin 2014).

Evaluating the set1-set3 trees with log-likelihood and as per the substitution models, LG, Dayhoff and BLOSUM62 are found to yield more accurate solutions through MUSCLE, MUSCLE and PROMALS3D alignments, against the respective worst trees formed by VT, LG and JTT models for the Kalign1 alignments (Pei and Grishin 2014). For the lowest tree length and internal tree length scores for sets1-3, Kalign1 consistently yields the best trees through PMB, PMB and BLOSUM62 matrices, against the respective worst trees formed by LG, DCMut and Dayhoff, and Dayhoff, DCMut and Dayhoff substitution models through MUSCLE, MUSCLE and PROMALS3D, and PROMALS3D, MUSCLE and MUSCLE1 alignments. However, screening the top-ranked 9 trees on the basis of the substitution models,

PMB, BLOSUM62, LG, Dayhoff, DCMut based solutions are found the best for 4, 2, 1 and 1 cases, although 7 KAlign1 and 2 MUSCLE alignments are only found to build these 9 trees.

Amongst the 18 evolutionary trees constructed for various sequence alignments and substitution models of the three sets, MUSCLE, MUSCLE1, Kalign1 and PROMALS3D protocols are observed to consistently yield more accurate trees for 4, 1, 11 and 2 cases for various substitution models viz. LG, PMB, Dayhoff, JTT and BLOSUM62. Hence, the alignment and substitution matrix protocols are statistically evaluated to find the consistently correct evaluation measure (Pei and Grishin 2014). Screening the best alignment protocols on the basis of the average statistics of the three measures, Kalign1, MUSCLE1 and MUSCLE protocols are found to yield the best trees for the 4, 3 and 2 cases, and MUSCLE1 yields the consistently correct evolutionary relationships for set3 for all scoring parameters. However, for sets 1 and 2, MUSCLE is found to yield the best log-likelihood trees and Kalign1 is found to build the trees with the lowest total tree length and internal tree length scores. Further, the standard deviation score for these three protocols is respectively found within the percentage range of 4.251 - 5.304. Conversely, the Kalign1 alignment is found to show the worst tree with the highest log-likelihood score for VT, LG and BLOSUM62 matrices for sets 1-3, and likewise is the case for MUSCLE and PROMALS3D for the tree-length based scores.

A similar analysis of the substitution measures likewise unfolds another complicity. It is observed that PMB, LG, Dayhoff and BLOSUM62 yield the best trees for 5, 2, 1 and 1 cases, with PMB model consistently being more accurate for the tree-length based measure. As per the total tree-length measure, the PMB matrix is found to be the best for all the three datasets (Philippe et al. 2011; Shalini Veerassamy 2003). For the internal tree length, BLOSUM62 and PMB matrix yield the best tree for set1 and sets 2 and 3, and for log-likelihood, LG and Dayhoff

models aid the construction of sets 1 and 3, and set 2 respectively. Further, the standard deviation against mean is found within the percentage range of 4.02 - 34.84 for the LG and PMB matrices for the log-likelihood and total tree length score for sets 3 and 2 respectively.

The analysis proves that both alignment protocol and substitution matrix are equally responsible for the construction of a biologically meaningful evolutionary tree (Jordan and Goldman 2012), and it indicates that the accuracy of an evolutionary protocol is highly dependent on the correct combination of these two protocols (Chang et al. 2014). Hence, the huge set of 270 (set 1-117, set 2-81, and set 3-72) phylogenetic trees are reanalyzed to decipher the top-ranked pairs of sequence alignment and the substitution models, and to find their consistently correct algorithmic set (Table V).

The analysis unleashes several intriguing features. MUSCLE alignment is found successful only for the smaller sequence datasets 1 and 2 (Deorowicz et al. 2016). The MUSCLE alignment, along with the LG and Dayhoff matrices, is found to yield the tree with the lowest log-likelihood score for 92.307% and 55.556% cases for the sets 1 and 2 respectively. For 270 constructed trees (92.307%), the MUSCLE alignment is found to work well with only the LG and Dayhoff substitution matrices for sequences shorter than 500 residues for the log-likelihood measure (Pearson 2013b), although for the distance based measure, the PMB matrix yields a more accurate tree through the -Kalign1 alignment (Nuin et al. 2006). However, the evolutionary trees constructed through it are maximally found better for the log-likelihood measure only, and its application is therefore not well-suited for drawing a robustly accurate phylogenetic relationship in view of the rapidly growing sequence data (Salichos et al. 2014). Further, PMB-Kalign1 protocol is found to yield the tree with the lowest tree length scores for 10 of the 13 set 1 trees (76.923%) and 7 of the 9 set 2 (77.778%) trees. However, for 712-

sequence set3, the PMB matrix yields the tree with the lowest tree-length score for the PROMALS3D and MUSCLE1 alignments for set2 and set3 respectively. Although, PMB substitution matrix is found to consistently construct more accurate trees with the lowest total tree length score for all the sequence datasets, it is found incompatible to handle the diverse datasets with varying size and length (Stamatakis 2014). A substantial divergence is observed in the tree-topology of the top-ranked tree against the sub-optimal solutions (Shukla et al. 2019). It implies that there is a specific biological upper limit for each alignment protocol to derive the meaningful results of the overlapping residues, and is in line with the earlier study (C S Troy 2001). Evaluating the trees for the location of various branches, it could be concluded that an accurate alignment is mandatory to draw the best-possible evolutionary relationship, as has been recently shown (Georgios A Pavlopoulos 2010). A significant disparity is found for all the trees for all the scoring parameters (Hall 2013), and selection of the best tree is further hereby shown to be difficult, especially for bigger dataset (Nguyen et al. 2015).

As an attempt to build more accurate evolutionary trees, the protocols including RaxML iteratively construct the alignments with a guide tree (Stamatakis 2014). Further, the “phylogeny-aware” tools like PRANK (Ari Löytynoja 2005) and PAGAN (Löytynoja et al. 2012) presume such guide-trees as the true trees to ensure that the constructed alignments are evolutionarily correct, and are thus termed as the post-tree analysis methods. Although the interdependence of MSA and inference of a phylogeny tree has been described in these methods, the best combination of alignment method and substitution matrix has not been greatly excavated for assessments.

Required strategic improvement over the existing protocols

The sequence alignments are usually constructed through several methodologies, and are substantially different for the number and placement of gaps, and as incorrect gaps lead to an inaccurate evolutionary tree, the alignments have always been iteratively constructed to find the best possible alignment and to derive the optimal phylogenetic solution (Blackburne and Whelan 2013; Fernández-Baca et al. 2004; Landan and Graur 2007; Sievers et al. 2013; T Heath Ogden 2006). Likewise, the substitution models are usually deployed to derive the evolutionary relationship among proteins, and different models have been tried for constructing a more accurate tree (Duchene et al. 2016; Hoff et al. 2016; Shapiro et al. 2006). However, for a sequence dataset, the phylogeny servers never iteratively optimize the evolutionary tree to save their processing time and computational resource. As demonstrated by our study, the biological significance of alignments/substitution models should be simultaneously tested through iterative assessments to construct the best evolutionary model for a sequence dataset. The study showcases a strategy to circumvent the potential pitfalls of the evolutionary algorithm. It will allow us consistently construct an evolutionary correct model for a sequence dataset and will subsequently improve the prediction accuracy of several different types of biological studies, be it a sequence characterization (Guder and Krishna 2019; Kalyani et al. 2019; Mallu et al. 2019; Satyanarayana et al. 2017; Somavarapu et al. 2017; Vemuri et al. 2019) or proteomic study (Jaswanthi et al. 2019; Rao et al. 2017; Samara Shekar Reddy et al. 2019; Srideepthi et al. 2017).

Strategic steps to assess and achieve biological accuracy of alignment and phylogeny

As recently shown by an evolutionary analysis of the foldIV transaminases with experimentally solved structures (Pavkov-Keller et al. 2016), all the four subfamilies, viz. 1-branched chain aminotransferases (BCATs), d-amino acid aminotransferases (D-ATAs), (R)-selective transaminases (RATA) and 4-amino-4-deoxychorismate lyases (ADCL), are found to be clustered into different clades (Francesco G. Mutti 2011; Percudani and Peracchi 2009). However, the phylogenetic protocol failed to accurately classify two functionally characterized RATA sequences CpuTA1 and MgiTA1, sharing a 49% sequence identity and containing conserved amino acids for both RATA and ADCL (Pavkov-Keller et al. 2016). It thus shows the need of functional investigation of the input sequences before constructing their biologically meaningful tree. Accurate establishment and quantification of homology through various substitution model/alignment algorithms is proven to be the key feature to accurately align the sequences and generate the meaningful phylogenetic trees. Moreover, evolutionary biologists should thus construct the consensus tree, as implemented by IQ-Tree server (Nguyen et al. 2015), and investigate the chance of shrouded negative factors through iterative assessments.

Conclusion

Evolutionary trees have long been deployed for several research methodologies. Here, we have focused on constructing an improved phylogenetic tree for a sequence dataset. The constructed set of 270 (set1-117, set2-81, set3-72) phylogenetic trees for the 36-sequence, 109-sequence and 712-sequence datasets illustrates that the accuracy of an evolutionary study is significantly determined by the compatibility of the sequence alignment and the residue substitution model,

and is not simply governed by the usually parameterized bootstrap and minimum correlation coefficient features. The best complementary set of biologically closest alignment and the substitution model is observed to yield the most-accurate evolutionary tree. Assessing the resultant phylogeny solutions on the basis of log-likelihood and minimal internal branch length, the study streamlines the methodology and opens avenues to design the robustly accurate phylogeny protocol. The study also lays a benchmark for the development of a protocol to synergistically deploy the assessment measures and consistently derive a more accurate evolutionary solution for all datasets.

References

- Alva V, Nam SZ, Soding J, Lupas AN (2016) The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis *Nucleic Acids Res* 44:W410-415 doi:10.1093/nar/gkw348
- Amrein BA, Runthala A, Kamerlin SCL (2019) In Silico-Directed Evolution Using CADEE Methods *Mol Biol* 1851:381-415 doi:10.1007/978-1-4939-8736-8_22
- Anil Kumar Jamithireddy AR, Balasubramaniam Gopal (2020) Evaluation of specificity determinants in mycobacterium tuberculosis σ /anti- σ factor interactions *Biochemical and Biophysical Research Communications* 521:900-906 doi:10.1016/j.bbrc.2019.10.198
- Arenas M (2015) Trends in substitution models of molecular evolution *Frontiers in Genetics* 6:319 doi:10.3389/fgene.2015.00319
- Ari Löytynoja NG (2005) An algorithm for progressive multiple alignment of sequences with insertions *Proceedings of the National Academy of Sciences* 102:6 doi:10.1073/pnas.0409137102
- Ashish R, Shibasish C (2013) Protein Structure Prediction: Are We There Yet? In: Tuan Pham LJ (ed) *Knowledge-Based Systems in Biomedicine*. vol SCI450. Springer, Berlin, Heidelberg, pp 79–115. doi:10.1007/978-3-642-33015-5_4
- Ashish Runthala AT, Dhiraj Sharma, Mahaveer Singh (2011) PHYLOGENETIC SELECTION GUIDED SACCHAROMYCES CEREVISIAE S288C GLUCOSE FERMENTATION MODELING *International Journal of Bioinformatics Research* 3:178-184
- Blackburne BP, Whelan S (2013) Class of multiple sequence alignment algorithm affects genomic analysis *Mol Biol Evol* 30:642-653 doi:10.1093/molbev/mss256
- Boc A, Diallo AB, Makarenkov V (2012) T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks *Nucleic Acids Research* 40:W573-579 doi:10.1093/nar/gks485
- BROOKS WK (1883) The phylogeny of the higher crustacea *Science* 2:4 doi:10.1126/science.ns-2.46.790
- Bruce Rannala ZY (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference *Journal of Molecular Evolution* 43:8 doi:10.1007/BF02338839
- C S Troy DEM, J F Bailey, D A Magee, R T Loftus, P Cunningham, A T Chamberlain, B C Sykes, D G Bradley (2001) Genetic evidence for Near-Eastern origins of European cattle *Nature* 410:4 doi:10.1038/35074088
- Chan L (1989) The Apolipoprotein Multigene Family: Structure, Expression, Evolution, and Molecular Genetics *Klinische Wochenschrift* 67:13

- Chang JM, Di Tommaso P, Notredame C (2014) TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction *Molecular Biology and Evolution* 31:1625-1637 doi:10.1093/molbev/msu117
- Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, Notredame C (2016) Multiple sequence alignment modeling: methods and applications *Briefings in Bioinformatics* 17:1009-1023 doi:10.1093/bib/bbv099
- Deorowicz S, Debudaj-Grabysz A, Gudys A (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families *Scientific Reports* 6:33964 doi:10.1038/srep33964
- Di Tommaso P et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension *Nucleic Acids Research* 39:W13-17 doi:10.1093/nar/gkr245
- Duchene S, Di Giallonardo F, Holmes EC (2016) Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales *Mol Biol Evol* 33:255-267 doi:10.1093/molbev/msv207
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Research* 32:1792-1797 doi:10.1093/nar/gkh340
- Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap *Evolution* 39:9 doi:10.1111/j.1558-5646.1985.tb00420.x
- Fernández-Baca D, Seppäläinen T, Slutzki G (2004) Parametric multiple sequence alignment and phylogeny construction *Journal of Discrete Algorithms* 2:271-287 doi:10.1016/s1570-8667(03)00078-9
- Francesco G, Mutti CSF, Desiree Pressnitz, Johann H. Sattler, Wolfgang Kroutil (2011) Stereoselectivity of Four (R)-Selective Transaminases for the Asymmetric Amination of Ketones *Advanced Synthesis & Catalysis* 353:7 doi:10.1002/adsc.201100558
- Georgios A Pavlopoulos TGS, Adriano Barbosa-Silva, Reinhard Schneider (2010) A reference guide for tree analysis and visualization *Biodata Mining* 3:1
- Guder DG, Krishna MSR (2019) Isolation and Characterization of Potential Cellulose Degrading Bacteria from Sheep Rumen *Journal of Pure and Applied Microbiology* 13:1831-1839 doi:10.22207/jpam.13.3.60
- Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference *Nucleic Acids Research* 33:W557-559 doi:10.1093/nar/gki352
- Hall BG (2013) Building phylogenetic trees from molecular data with MEGA *Mol Biol Evol* 30:1229-1235 doi:10.1093/molbev/mst012
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation *Molecular Biology and Evolution* 35:518-522 doi:10.1093/molbev/msx281
- Hoff M, Orf S, Riehm B, Darriba D, Stamatakis A (2016) Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics* 17:143 doi:10.1186/s12859-016-0985-x
- Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies *Molecular Biology and Evolution* 23:254-267 doi:10.1093/molbev/msj030
- Jaswanthi N, Krishna MSR, Sahitya UL, Suneetha P (2019) Apoplast proteomic analysis reveals drought stress-responsive protein datasets in chilli (*Capsicum annum* L.) *Data in Brief* 25 doi:10.1016/j.dib.2019.104041
- Jordan G, Goldman N (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection *Mol Biol Evol* 29:1125-1139 doi:10.1093/molbev/msr272
- Kalyani BS, Krishna PS, Sreenivasulu K (2019) Screening and identification of novel isolate *Streptomyces* sp., NLKPB45 from Nellore coastal region for its biomedical applications *Saudi J Biol Sci* 26:1655-1660 doi:10.1016/j.sjbs.2018.08.027

- Kamjula V, Kanneganti A, Metla R, Nidamanuri K, Idupulapati S, Runthala A (2020) Decoding the vital segments in human ATP-dependent RNA helicase *Bioinformatics* 16:160-170 doi:10.6026/97320630016160
- Kazutaka Katoh KM, Kei-ichi Kuma, Takashi Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform *Nucleic Acids Research* 30:8
- Keeling PJ, Fast NM, McFadden GI (1998) Evolutionary relationship between translation initiation factor eIF-2gamma and selenocysteine-specific elongation factor SELB: change of function in translation factors *Journal of Molecular Evolution* 47:649-655 doi:10.1007/pl00006422
- Kemena C, Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era *Bioinformatics* 25:2455-2465 doi:10.1093/bioinformatics/btp452
- Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution *Journal of Royal Society Interface* 5:151-170 doi:10.1098/rsif.2007.1047
- Kosiol C, Bofkin L, Whelan S (2006) Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome *Journal of Biomedical Informatics* 39:51-61 doi:10.1016/j.jbi.2005.08.003
- Kosiol C, Goldman N (2005) Different versions of the Dayhoff rate matrix *Molecular Biology and Evolution* 22:193-199 doi:10.1093/molbev/msi005
- Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments *Mol Biol Evol* 24:1380-1383 doi:10.1093/molbev/msm060
- Lassmann T, Sonnhammer EL (2005) Kalign--an accurate and fast multiple sequence alignment algorithm *BMC Bioinformatics* 6:298 doi:10.1186/1471-2105-6-298
- Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix *Molecular Biology and Evolution* 25:1307-1320 doi:10.1093/molbev/msn067
- Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists *Nucleic Acids Research* 47:W260-W265 doi:10.1093/nar/gkz303
- Loytynoja A, Vilella AJ, Goldman N (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm *Bioinformatics* 28:1684-1691 doi:10.1093/bioinformatics/bts198
- Mallu MR, Dronavalli N, Nannapaneni S, Kamarapu A, Vemula S (2019) Detection and Characterization of CryIac in BT Cotton Hybrids of MECH 162 and RCH2 *Research Journal of Pharmacy and Technology* 12 doi:10.5958/0974-360x.2019.01015.1
- Minh BQ, Nguyen MA, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap *Mol Biol Evol* 30:1188-1195 doi:10.1093/molbev/mst024
- Nabhan AR, Sarkar IN (2012) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy *Brief Bioinform* 13:122-134 doi:10.1093/bib/bbr014
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies *Molecular Biology and Evolution* 32:268-274 doi:10.1093/molbev/msu300
- Nikoh N, Hayase N, Iwabe N, Kuma K, Miyata T (1994) Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi, inferred from 23 different protein species *Molecular Biology and Evolution* 11:762-768 doi:10.1093/oxfordjournals.molbev.a040156
- Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins *BMC Bioinformatics* 7:471 doi:10.1186/1471-2105-7-471
- Nunes LA, Turvey ST, Rosindell J (2015) The price of conserving avian phylogenetic diversity: a global prioritization approach *Philosophical Transactions of the Royal Society B* 370:20140004 doi:10.1098/rstb.2014.0004
- Parks SL, Goldman N (2014) Maximum likelihood inference of small trees in the presence of long branches *Systematic Biology* 63:798-811 doi:10.1093/sysbio/syu044

- Pavkov-Keller T et al. (2016) Discovery and structural characterisation of new fold type IV-transaminases exemplify the diversity of this enzyme fold *Scientific Reports* 6 doi:10.1038/srep38183
- Pearson WR (2013a) An Introduction to Sequence Similarity (“Homology”) Searching *Current Protocols in Bioinformatics* 42:3.1.1-3.1.8 doi:10.1002/0471250953.bi0301s42
- Pearson WR (2013b) Selecting the Right Similarity-Scoring Matrix *Curr Protoc Bioinformatics* 43:3 5 1-3 5 9 doi:10.1002/0471250953.bi0305s43
- Pei J, Grishin NV (2014) PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information *Methods in Molecular Biology* 1079:263-271 doi:10.1007/978-1-62703-646-7_17
- Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments *Nucleic Acids Research* 36:2295-2300 doi:10.1093/nar/gkn072
- Percudani R, Peracchi A (2009) The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families *BMC Bioinformatics* 10:273 doi:10.1186/1471-2105-10-273
- Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D (2011) Resolving difficult phylogenetic questions: why more sequences are not enough *PLoS Biology* 9:e1000602 doi:10.1371/journal.pbio.1000602
- Phulara SC, Rajput VS, Mazumdar B, Runthala A (2020) Metabolic and Enzyme Engineering for the Microbial Production of Anticancer Terpenoids. In: 'Essentials of Cancer Genomic, Computational Approaches and Precision Medicine. pp 237-259. doi:10.1007/978-981-15-1067-0_10
- Rao DK et al. (2017) Distribution of CYP2C8 and CYP2C9 amino acid substitution alleles in South Indian diabetes patients: A genotypic and computational protein phenotype study *Clin Exp Pharmacol Physiol* 44:1171-1179 doi:10.1111/1440-1681.12810
- Runthala A (2012) Protein structure prediction: challenging targets for CASP10 *Journal of Biomolecular Structure and Dynamics* 30:607-615 doi:10.1080/07391102.2012.687526
- Runthala A (2015) Refinement and Improvement of Template Based Protein Modelling Algorithms. BITS Pilani
- Runthala A (2020) Probabilistic divergence of a TBM methodology from the ideal protocol *BiorXiv* doi:10.1101/2020.07.05.160937
- Runthala A, Chowdhury S Iterative Optimal TM_Score and Z_Score Guided Sampling Significantly Improves Model Topology. In: International MultiConference of Engineers and Computer Scientists, Hong Kong, March 12 - 14 2014. International Association of Engineers. doi:10.13140/2.1.4584.5443
- Runthala A, Chowdhury S (2016) Unsolved Problems of Ambient Computationally Intelligent TBM Algorithms. In: Hybrid Soft Computing Approaches. Studies in Computational Intelligence. pp 75-105. doi:10.1007/978-81-322-2544-7_3
- Runthala A, Chowdhury S (2019) Refined template selection and combination algorithm significantly improves template-based modeling accuracy *J Bioinform Comput Biol* 17:1950006 doi:10.1142/S0219720019500069
- Sagulenko P, Puller V, Neher RA (2018) TreeTime: Maximum-likelihood phylodynamic analysis *Virus Evolution* 4:vex042 doi:10.1093/ve/vex042
- Salichos L, Stamatakis A, Rokas A (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees *Molecular Biology and Evolution* 31:1261-1271 doi:10.1093/molbev/msu061
- Samara Shekar Reddy S, Singh B, John Peter A, Venkateswar Rao T (2019) Genetic transformation of indica rice varieties involving Am-SOD gene for improved abiotic stress tolerance *Saudi J Biol Sci* 26:294-300 doi:10.1016/j.sjbs.2017.06.009
- Satyanarayana SD, Msr K, Pindi PK (2017) A strategic approach for isolation and identification of plant growth promoting rhizobial strains from bhadrachalam forest area with respect to soybean (*Glycine max*) *International Journal of Pharma and Bio Sciences* 8 doi:10.22376/ijpbs.2017.8.2.b401-408

- Shalini Veerassamy AS, Elisabeth RM Tillier (2003) A Transition Probability Model for Amino Acid Substitutions from Blocks Journal of Computational Biology 10:14 doi:10.1089/106652703322756195
- Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences Mol Biol Evol 23:7-9 doi:10.1093/molbev/msj021
- Sharma A, Rani S, Goel M (2018) Navigating the structure-function-evolutionary relationship of CsaA chaperone in archaea Crit Rev Microbiol 44:274-289 doi:10.1080/1040841X.2017.1357535
- Shukla AK, Tripathi D, Reddy BR, Chandramohan D (2019) A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges Evolutionary Intelligence 13:309-329 doi:10.1007/s12065-019-00306-6
- Sievers F, Dineen D, Wilm A, Higgins DG (2013) Making automated multiple alignments of very large numbers of protein sequences Bioinformatics 29:989-995 doi:10.1093/bioinformatics/btt093
- Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences Protein Science 27:135-145 doi:10.1002/pro.3290
- Sievers F et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega Molecular Systems Biology 7:539 doi:10.1038/msb.2011.75
- Somavarapu S, Vemula S, Reddy IB (2017) Extraction, purification and characterization of a novel cysteine protease from the latex of plant Vallaris solanacea Journal of Plant Biochemistry and Biotechnology doi:10.1007/s13562-017-0429-3
- Srideepthi R, Lakshmisahitya U, Peddakasi D, Suneetha P, Krishna M (2017) Morphological, Pathological and Molecular Diversity of Colletotrichum capsici inciting Fruit Rot in Chilli (Capsicum annum L.) Research Journal of Biotechnology 12:8
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies Bioinformatics 30:1312-1313 doi:10.1093/bioinformatics/btu033
- T Heath Ogden MSR (2006) Multiple sequence alignment accuracy and phylogenetic inference Systematic Biology 55:15 doi:10.1080/10635150500541730
- T. Lanis`nik Riz`ner JS, J. Adamski (2001) 17b-hydroxysteroid dehydrogenase from the fungus Cochliobolus lunatus: structural and functional aspects Chemico-Biological Interactions 130-132:11
- Vemuri PK, Yarlagadda DV, Tatineni J (2019) Characterization of Sesamum indicum proteins and its immunogenic activity Drug Invention Today 11:5
- Winona C. Barker LKK, Margaret O. Dayhoff (1978) A Comprehensive Examination of Protein Sequences for Evidence of Internal Gene Duplication Journal of Molecular Evolution 10:17
- Xuan Y et al. (2019) Molecular phylogeny of mulberries reconstructed from ITS and two cpDNA sequences PeerJ 7:e8158 doi:10.7717/peerj.8158