Tel-Aviv University
The George S. Wise Faculty of Life Sciences
Graduate School

Towards realistic sequence simulations for the reconstruction of mega phylogenies

Thesis submitted towards the M.Sc. degree in Bioinformatics

in Tel-Aviv University

by

Nomi Hadar

The research was performed in the Department

of Molecular Biology and Ecology of Plants

Under the supervision of

Prof. Itay Mayrose

June 2018

תודות

"אתה חונן לאדם דעת חננו מאתך חכמה בינה ודעת" (תפילת שמונה-עשרה)

ראשית תודה לבורא-עולם על שכיוון אותי אל המעבדה שהייתה לי כבית שני במשך שנתיים טובות ועשירות, וחנן אותי ביכולת להביא לידי גמר עבודה זו.

פרופ' איתי מירוז, מנחה התזה שלי, הוא חלומו של כל סטודנט לתואר שני. מעבר למקצועיותו הבלתי מוטלת בספק, הוא ניחן בסבלנות ובאנושיות מעוררות השראה. מהרגע הראשון בו נכנסתי לממלכתו, זרה בעיר הגדולה וחסרת ניסיון קודם במחקר, הוא קיבל אותי במאור פנים וגילה התעניינות שלא התפוגגה עם השנים. אחת מתכונותיו המופלאות ביותר היא הצלילה לפרטים הקטנים ביותר - עם במחקר, החל בנוהגו לפרוט כל משימה לשלבים מוגדרים ונהירים וכלה במעורבותו באחרון הפרמטרים הנדרשים להרצה מסוימת, ואם בדאגה לנוחיותי, מסביבת העבודה ועד לדרכי ההגעה. תודה על ההכוונה, העידוד והתמיכה ברגעי משבר. תודה על ההתחשבות באורח-חיי. תודה על הדלת הפתוחה והזמינות התמידית. תודה על היכולת לשבח נקודות טובות ולחזק את הדורשות חיזוק. תודה על אורך-הרוח לפשט ולבאר עד ליישור המצח. תודה על תיקון אינספור הטיוטות שהיו גלגוליהן של עבודה זו. תודה על הזכות לעשות את צעדי הראשונים במחקר במחיצתך.

חברי למעבדה בעבר ובהווה הם הסיבה העיקרית בשלה קמתי בשמחה כל יום. ברצוני להודות לכל אחד ואחד מהם על מי שהם. תודה מיוחדת לשירן עבדי על היותה עמוד התווך של המעבדה והתגלמות הביטוי "ראש גדול". תודה על אינספור הבעיות שפיצחת עבורי, על שלל העצות לחיים שנידבת לי ועל היותך חברה כה נהדרת. תודה לדנה עזורי על היותה חברה אהובה, תומכת וקשובה שיודעת לקרוא אותי במבט אחד. תודה לאנה רייס שליוותה אותי בשלבי המחקר המוקדמים וחלקה איתי מניסיונה. תודה לקרן חלבי על שיחות פילוסופיות עמוקות ומרתקות. תודה למיכל דרורי על ההנאה לעבוד לצידה על אף חילוקי הדעות. תודה לענת שפיר על שמדגימה לי יום-יום כיצד להשמיע דברים בנחת. תודה לאילת סלמן על הלבביות בה הייתה מקבלת את פני בכל בוקר ושומעת את תלונותיי על הפקקים. תודה לגל היימס על דיונים מעניינים בארוחות-הצהריים. תודה לליאור גליק על שחזר למעבדה לא מזמן והבנתי מדוע שמחו בחזרתו. תודה להילה צפדיה היקרה על השמחה בי בכל מפגש ועל השעות שבילינו יחד.

ברצוני להודות גם לפרופ' טל פופקו על שהקדיש לי מזמנו וליווה את המחקר בשלביו האחרונים, וכן למי שהיו הדוקטורנטים שלו: ד"ר אלי לוי-קרין לד"ר חיים אשכנזי, שניהם חוקרים מבריקים ומרתקים, שהתוכנות שפיתחו הן חלק בלתי נפרד מהעבודה הזאת. אלי, תודה שהראית לי מהי התלהבות מדעית ומהי מופת של הרצאה. חיים, תודה על הזמינות לשאלות במייל ופנים-אל-פנים ועל הידע שחלקת עמי.

לבסוף, ברצוני להודות להורי, ישראל ויהודית, אשר נטעו בי את האהבה לדעת, ואפשרו לי ללמוד ולהתפתח בראש שקט במשך כל שנות נעורי ובגרותי. צערי היחיד הוא על כך שעבודה זו לא תהייה קריאה עבורם. כולי מקווה שביום מן הימים אכתוב משהו שיוכלו לעיין ולשמוח בו.

תזה זו מוקדשת לאחי הקטן והאהוב, אריאל, שהאכפתיות והאהבה שלו משכיחות ממני לעיתים את שלוש-עשרה השנים העומדות בינינו.

Abstract

The reconstruction of large phylogenies encompassing thousands of species enables us a wide view on evolutionary processes and serve as the starting point for the study of numerous evolutionary phenomena. Methods for phylogeny reconstruction are based on an input multiple sequence alignment (MSA), encompassing multiple loci. The extent to which the inferred phylogeny will reflect the true history is thus directly dependent on the accuracy of the alignment. However, the task of sequence alignment becomes more challenging with the amount of divergence among the analyzed sequences. Thus, the reconstruction of large alignments consisting of broadly sampled sequences might lead to poor alignments and inaccurate trees. In this study, I initially aimed to evaluate the effect of alignment complexity on the accuracy of the inferred phylogenies through sequence simulations. Particularly, I examined the hypothesis that the more complex the alignment is the less accurate the inferred phylogeny will be. Such effect was indeed observed; however, it was lower than expected. I suspect that sequence simulations are not realistic enough, rendering them non-informative to study the accuracy of mega phylogenies. My research thus focused on identifying the shortcomings of current simulation approaches and on the pathways by which sequence simulations could be improved. First, I found that available methodologies for inferring the parameters controlling indel dynamics fail to produce alignments that resemble the reference alignments. This required me to develop a new procedure to infer these parameters. Second, using alignment reliability measures, I found that the simulated MSAs differ drastically in their alignment complexity from real MSA, i.e., the complexity of the simulated MSAs was much lower than that of the real one. Third, I found that the inclusion of short, partial, sequences had a major effect on alignment complexity. Fourth, I found that the power-law distribution that is commonly used to describe the length-distribution of indel events, may not be the most suitable one. In summary, my conclusion is that the sequence simulation field is in its early stages, and much study is still required to achive more realistic simulations. Until then, the conclusions that are drawn based on current simulations should be taken with great caution.

Table of contents

1. Introduction

1.1. Large phylogenies

Inferring the evolutionary relationships among species is one of the oldest and most basic tasks of evolutionary research. The prototype phylogenetic tree, or the "tree of life", was conceived by Charles Darwin in his seminal book "The Origin of Species", and its inference has been approached by generations of biologists ever since. The rapid accumulation of sequence data in recent years has enabled the reconstruction of large phylogenies in a scale that could be only imagined a decade ago. Over the past few years, great efforts to reconstruct large phylogenies encompassing thousands of species have been made. For example, recently published phylogenies includes those of seed plants (Smith and Brown, 2018) with 280,000 species, birds (Jetz *et al.*, 2012) with 9,990 species, and Squamata (Pyron, Burbrink and Wiens, 2013) with 4,000 species. These advances, however, come with considerable gaps of knowledge regarding the efficacy and accuracy of current approaches to deal with such massive amounts of data.

1.2. Sequence alignment

Following the assembly of the sequence data, aligning a set of modern orthologous sequences is often the first step of phylogeny reconstruction (Fig 1). Two or more sequences are orthologous if they share a common ancestor, i.e., they descendent from the same ancestral sequence. Through the course of evolution, each sequence changes independently with respect to the common ancestral sequence. Such sequence changes include events of substitution, insertion, and deletion (indel) (Fig 2), yielding modified, possibly longer or shorter sequences. A multiple sequence alignment (MSA), is a tabular representation of the

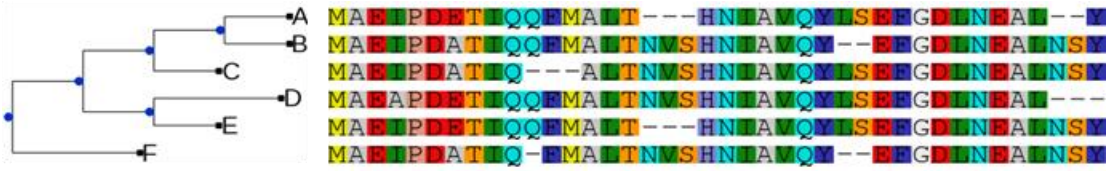
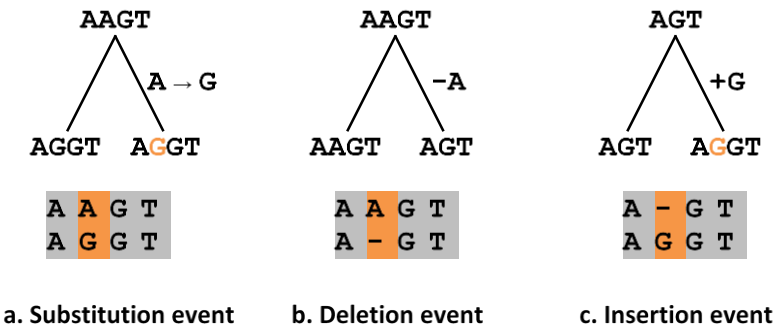


Fig 1. An example of an MSA and its associated phylogenetic tree. The evolutionary relationship among six taxa (A-F) is illustrated using the phylogeny at the left. The aligned sequence data of each species is presented to the right of each tip.



a. Substitution event **b. Deletion event** **c. Insertion event**

Fig 2. Possibly sequence changes and their representation in a multiple sequence alignment.

resulting position-specific homology relationships among the sequences. Each alignment row is an ortholog sequence and each alignment column contains homologous characters with respect to the common ancestor (Kumar and Filipski, 2007). Positions that have no homologous characters as a result on an indel event are represented with a gap character, usually marked with '-'.

1.3. Approaches for large phylogenies reconstruction

There are two major approaches to accomplish the task of inferring mega phylogenies: super-trees and super-matrices (Smith, Beaulieu and Donoghue, 2009) (Fig 3). In the super-tree approach, multiple data sets are analyzed separately, and then the trees derived from these independent analyses are combined to produce a single large phylogeny. Alternatively, the super-matrix approach involves a two-step procedure: first, the multiple sequence alignments of multiple loci, each containing all or a fraction of all relevant species, is computed. The alignments are then concatenated into a single sequence matrix, potentially spanning hundreds of genomic loci. Second, given the super MSA, the phylogeny is reconstructed using any tree building procedure, while considering the differences in evolutionary dynamics among each loci. The super-matrix method is considered superior to the super-tree method (de Queiroz and Gatesy, 2007) and my research focuses on this approach.

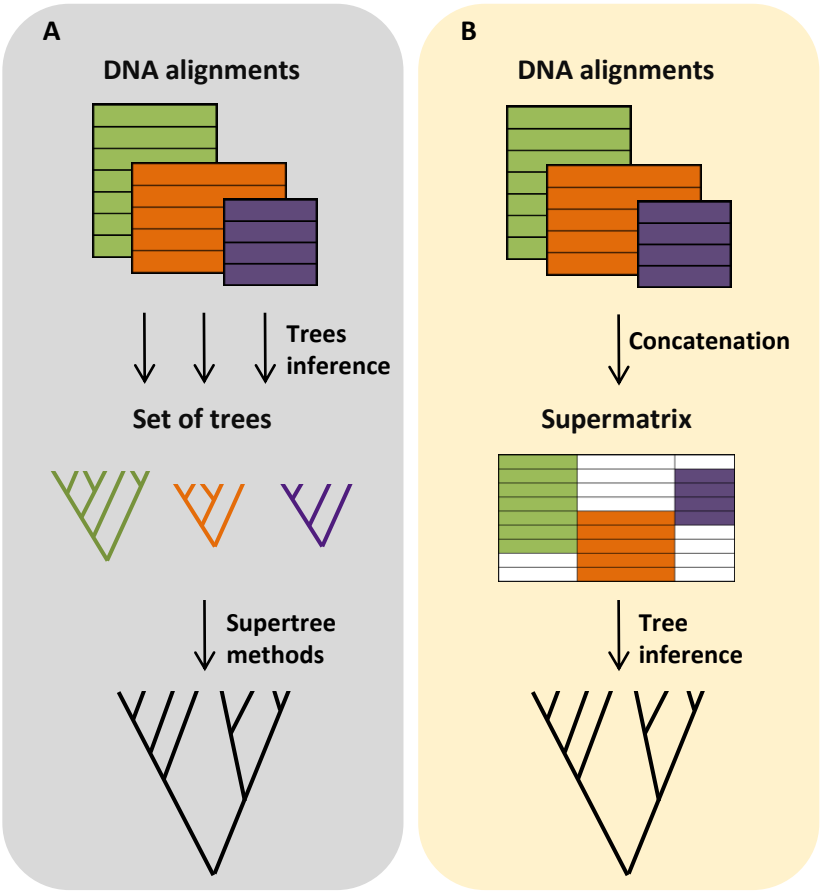


Fig 3. Approaches for mega-phylogeny reconstruction. (A) The supertree method first reconstructs a set of small trees and then assembles the supertree for all species. **(B)** The supermatrix method concatenates sequences from multiple loci into a data supermatrix for a single inferred phylogeny.

1.4. PHLAWD – a tool for construction phylogenetic super-matrices

PHLAWD (Smith, Beaulieu and Donoghue, 2009) is a software for the construction super-matrices, with the goal to address the problem of aligning divergent sequences. It retrieves GenBank sequences of a clade of interest and subsequently constructs their multiple alignment. The alignment construction is done by employing profile alignments to combine alignments of orthologous gene regions (that would otherwise be poorly aligned if performed across the entire group). To retrieve the sequences the user has to identify the gene regions of interest by presenting actual examples of the gene region and the breadth of molecular diversity of that gene within the clade of interest.

A profile alignment is an algorithmic approach in which the sequences are separated into subgroups by some criteria, and then the alignments of each group are combined to produce

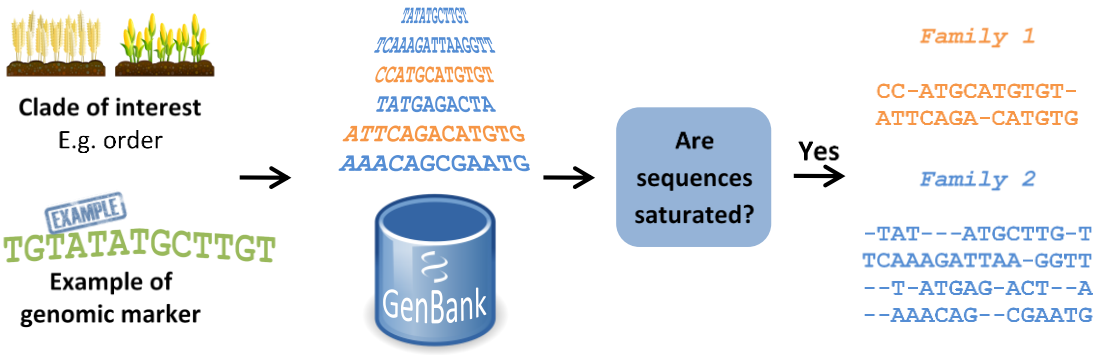


Fig 4. Scheme of PHLAWD pipeline. Starting from a clade of interest and a representative of the target genomic marker, PHLAWD extracts from GenBank sequences of species within the specified clade with similarity to the given representative sequence. The level of the saturation in the group of sequences is tested, and if the group is found to be saturated - then the group is broken up into less inclusive group (here the order group in broken into families), and each group is aligned separately. The resulted alignments are then combined into a single alignment.

a single alignment. In PHLAWD, sequences are separated into subgroups of aligned sequences based on the degree of sequence saturation[1]. For example, if the most inclusive group of sequences is deemed saturated, then the group is broken up into less inclusive groups using the next level in the taxonomic hierarchy. In a Linnaean taxonomic system, if an "order" is found to be saturated, it would be broken into "families". Each smaller subset of sequences is then re-aligned and the degree of saturation reassessed (however, the taxonomic groups used in this procedure need not correspond to ranks in the Linnaean hierarchy, but should simply be hierarchically nested, as in the NCBI taxonomy). This process continues iteratively to less inclusive groups until sequences no longer appear saturated and these alignments are then stored. After every sequence has been either placed in an alignment or placed at a "singleton" the individual alignments are then profiled to a larger alignment (Fig 4).

1.5. Phylogenies accuracy

Importantly, the reconstructed phylogenies do not necessarily provide an accurate representation of the true evolutionary history of the included taxa. In fact, these phylogenies should be considered scientific hypotheses, subject to falsification by further study.

[1] Genetic saturation is the reduced appearance, which occurs over time, of sequence divergence rate that results from reverse mutations, homoplasies, and other multiple changes occurring at single sites along two lineages

Specifically, the availability of large number of sequences is not necessarily associated with more accurate phylogenies. The examination of recently-inferred mega phylogenies revealed that the resulting trees are often of poor quality (Hinchliff and Smith, 2014). A demonstration of this last claim will also be investigated in the Results section.

1.5.1. Accuracy of sequence alignment methods

Multiple sequence alignment is a standard stage in most phylogenetic reconstruction methods (Edgar and Batzoglou, 2006; Notredame, 2007). Alignment-free approaches exist (Vinga and Almeida, 2003) but are not discussed here. It is also a crucial step in most of the advanced homology-based sequence analyses such as identifying conserved regions and finding molecular function. However, the most rigorous methods for phylogenetic reconstruction will not be able to reconstruct accurate evolutionary relationships if they are applied to unreliable multiple sequence alignments, therefore great care should be taken to ensure the accuracy of the alignments.

MSA computation stands at a cross-road between computation and biology. The computational issue is the complexity of an exact MSA computation: the problem of aligning a single pair of sequences has an optimal solution (Needleman and Wunsch, 1970), but aligning several sequences into a multiple sequence alignment is known to be NP-complete (i.e., a problem that cannot be solved in polynomial time in any known way) (Wang and Jiang, 1994). MSA computation therefore depends on approximate algorithms or heuristics (Durbin *et al.*, 1998), and most commonly involves the iterative computation of many smaller alignments. Over the last 30 years, more than 100 MSA methods have been published (Kemena and Notredame, 2009), based on all kind of heuristics, including the most widely used progressive method (Hogeweg and Hesper, 1984), implemented by many MSA packages such as MAFFT (Katoh and Standley, 2013) and ClustalW (Thompson, Higgins and Gibson, 1994). Progressive alignment sequentially builds up an MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related. The primary problem of the progressive technique is that errors made at any stage accumulate and propagate to the final result.

Moreover, heuristic alignment algorithms often use a "guide tree" to dictate the order of alignment steps. Ideally, the guide tree should reflect the phylogenetic relationships among the aligned sequences. However, the true phylogenetic relationship is generally unknown and the guide tree is reconstructed through a fast and inaccurate procedure. Furthermore, due to the heuristic nature of alignment procedures each of the various heuristic stages of any alignment algorithm has the potential of introducing errors. As a result, different algorithms often produce different MSAs given the same input set of unaligned sequences (Blackburne and Whelan, 2012).

The biological issue surrounding MSAs is even more complex. Given a set of sequences, we do not know how to estimate homology in a way that will guaranty the evolutionary correctness of an alignment (in fact, being able to do it implies having solved the phylogeny). In practice, multiple alignments are computed by maximizing identity, in the hope that this criterion will be sufficiently informative to yield models usable for most type of biological inference. The objective function that is being maximized is usually defined with an affined substitution matrix and a gap penalty scheme. The degree to which sequences differ is qualitatively related to their evolutionary distance from one another. Roughly speaking, high sequence identity suggests that the sequences have diverged from each other relative recently, while low identity suggests that the divergence is more ancient. However, even this simple strategy for inferring homogeneity requires a priori models about the evolution of the sequences that are being compared. The parameter values assumed by these models, such as the substitution matrix and gap penalties, are somewhat arbitrary. For example, reference substitution matrices for protein alignments are often used without the verification whether they are representative of the sequences being aligned. Moreover, the scoring system is not consensual between applications, and many reports have shown that small changes in the input parameters can greatly affect the resulting alignment (Wong, Suchard and Huelsenbeck, 2008).

The task of sequence alignment becomes more challenging as sequences become more divergent – either when it comes to large matrices of broadly sampled sequences or smaller matrices with rapidly evolving loci (Hickson, Simon and Perrey, 2000; Van Walle, Lasters and Wyns, 2004; Huang, Umbach and Li, 2005; Xia, 2016). Rapidly evolving regions are good for

solving recent events, but they are hard to align when the sequences are too far apart, resulting in poor alignments. The reason for the poor results is that the guide-trees built during the alignment estimation procedures are often based on raw pairwise (uncorrected) or model-corrected (e.g. Jukes-Cantor (1969)). These methods are susceptible to problems with saturation - multiple mutations at the same site for the same organism - and are therefore much less accurate for large and broadly sampled alignments that are likely to contain very distantly related sequences. Slowly evolving regions, in contrast, are less challenging to align, but due to insufficient signal can only be used for solving ancient divergence events, and not recent ones.

Taking together, the computation of multiple sequence alignment is a complicated statistical estimation problem, in which alignment uncertainty originates from both the difficulty of defining homologs and computational limitations of current evolutionary models and alignment methodologies. Alignment errors have been shown to affect various downstream analyses, leading, for example, to erroneous phylogenetic tree reconstruction (Crowder *et al.*, 2006; Soria-Carrasco *et al.*, 2007; Liu *et al.*, 2011).

1.5.2. Accuracy of phylogenetic reconstruction methods

Phylogenetic reconstruction methods can be divided into two types: those that proceed algorithmically (e.g., Neighbor-joining; Saitou and Nei, 1987) and those that are based on optimality criteria. Among the latter, probabilistic methods – either maximum likelihood or Bayesian are the most popular. Both are likelihood-based and allow for the implementation of multiple data partitions that can each be analyzed with its best-fit evolutionary model.

Maximum likelihood (ML) inference attempts to identify the phylogeny (including the topology and the branch lengths) that explains the evolution of a set of aligned sequences under a given model of evolution with the greatest likelihood (Felsenstein, 1981). Since ML reconstruction is NP-hard (Chor and Tuller, 2005), heuristic strategies are used, and therefore the optimal solution is not guaranteed. Several popular tools are available, such as RAxML (Stamatakis and Aberer, 2013), ExaML (Kozlov, Aberer and Stamatakis, 2015), and PhyML (Guindon and Gascuel, 2003) that implement ML phylogeny reconstruction efficiently and can handle large datasets of sequences. Bayesian inference combines the prior probability of a

phylogeny with the likelihood to produce a posterior probability distribution of trees, which can be interpreted as the probability that the tree is correct, given the specified model and priors (Huelsenbeck and Ronquist, 2001). Baysian methods are computationally demanding and their usability to reconstruct mega phylogenies has not been attempted.

As the MSAs reconstruction methods, all models used by the phylogenetic reconstruction methods incorporate assumptions about the evolutionary processes which are, by necessity, a great simplification. In addition, the event with the highest likelihood is only inferred to be so under the model assumption and not necessarily has occurred. Nevertheless, the accuracy of the inferred phylogeny is often thought to be more dependent upon the correctness of the sequence alignment rather than on the method of reconstruction (Ogden and Rosenberg, 2006).

1.5.3. Methods for assessing accuracy of MSAs and phylogenetic reconstruction

The true evolutionary relationships among the analyzed sequence is usually unknown, making it difficult to examine the relative accuracy of the methods for MSAs and phylogenetic reconstruction. Here I briefly discuss several methods for this task. The accuracy assessment of MSA programs is often done by employing manually (or semi automatically) curated databases of structure-based alignments. Structural alignment attempts to establish homology between two or more proteins based on their tertiary structures. An example of such database is BAliBASE (Thompson *et al.*, 2005), which is widely used as a reference standard. Alignment databases provide a source of reference alignments to evaluate the performance of different programs, but they also present several disadvantages. The databases are curated manually, with potential arbitrary and uneven biases resulting from human intervention. The sets of alignments are still rather small and may not represent the complete range of scenarios of protein evolution. Another drawback of the use of reference alignment databases is that algorithms can potentially be tuned to the alignments present solely in these data sets (i.e., overfitting).

Another approach to assess the accuracy of phylogenetic programs is the use of experimental evolution to obtain experimental support for phylogeny inference methods. This process, termed experimental phylogeny (Hillis *et al.*, 1992; Bull *et al.*, 1993) is the study of

evolutionary processes occurring in experimental populations in response to conditions imposed by the researcher. Since the diversification of the population is controlled, the complete true phylogeny is known. This known phylogeny enables testing the effectiveness of methods for inferring phylogeny by comparing the inferred evolutionary history against the true one. While microorganisms can be manipulated in the laboratory through thousands of generations per year, and mutation rates can be elevated through the use of mutagens, experiments with animals and plants have limitations of small population sizes, limited timescales, little genetic differentiation and the simplified nature of laboratory environments. Experimental studies of phylogenies are therefore feasible mainly with microorganism. Additionally, results obtained using experimental evolution techniques revealed that most methods reconstruct the same, true, phylogeny, indicating that this procedure is still rather simplistic compared to those that occur in nature.

Another way of assessing phylogenic accuracy is using taxonomy. Indeed, since the beginnings of phylogenetics, researchers have used a combination of phylogenetic inference and taxonomic knowledge to understand evolutionary relationships. Taxonomic classifications are often used to diagnose problems with phylogenetic inferences, for example by assessing the monophyly of specified ranks (i.e. families and orders), and conversely, phylogeny is used to bring taxonomists up to date with recent inferences and to identify misclassification (Hinchliff and Smith, 2014; Zapata *et al.*, 2015). NCBI Taxonomy (Sayers, 2009) and The Plant List (http://www.theplantlist.org/) are few examples of broadly used taxonomy databases.

Bootstrap is another method for evaluating the quality of a phylogeny. Unlike the other above-mentioned methods, it does not involve an additional source of information other than the data at hand. Bootstrapping is a general technique of constructing confidence measures based on the variability observed in the data. The assumption is that the variability in the observed data (several molecular markers) can approximate the true variation in the full data (the whole genome). For each split in the phylogeny, bootstrap support values for the phylogenetic reconstruction (Felsenstein, 1985) are obtained by sampling alignment columns with replacement to create a large number of pseudo-replicate data sets that are each of the same size as the original data set. The reconstruction analysis is then executed on each pseudo-replicate and the resulting trees are used to generate frequencies indicating how often a given

split of the original tree occurs in the set of pseudo-replicates. High bootstrap values are indicative of splits that are robust to noise in the data, while low bootstrap values suggest low support. Importantly, bootstrapping is not a measure of reliability of the phylogeny or its reconstruction method. Rather, bootstrapping is an ad-hoc procedure for determining whether the overall phylogenetic signal is consistent within the data at hand and, therefore, trustworthy. Moreover, bootstrapping assumes independence among the analyzed characters, an assumption that is made for computational tractability, but is shown to be overly simplistic.

The lack of known true MSAs and phylogenies led to the development of simulation techniques. These methods generate artificial datasets by simulating sequences along a given (or, true) phylogeny. The produced data is then used to reconstruct a phylogeny by one or more inference tools. The inferred and true trees can then be compared, thereby assessing the accuracy of the reconstruction procedure. The concept of sequence simulations is explained in more detail in the next section.

1.6. Modeling sequence evolution

Models are a set of definitions and assumptions designed to describe the nature and behavior of a system. When the model is simple enough, it may be possible to use mathematical rules to obtain an analytic solution for the question of interest. However, the internal complexity of real-world systems can rarely be captured by such simple models. In order to study elaborated models, we use simulations. A simulation is a computational approach that aims to mimic the behavior of the system under the model's assumptions. Simulations are a powerful tool to investigate the accuracy and robustness of an extraordinary range of systems (Law, Kelton and Kelton, 1991), as they allow full control of every aspect of the system. Another important advantage of simulations is the ability to fast-forward a process that in real life could have taken an immensely long of time. Simulations play a major role in biological studies, where processes can get extremely complicated, non-deterministic and endure for many generations. They are used for predicting living system behaviors (Worobey, Han and Rambaut, 2014), for testing and comparing computational methods (Loew and Schaff, 2001; De Jong, 2002; Slepchenko *et al.*, 2002), and are commonly applied in the field of molecular evolution, where the simulated data is a set of DNA, RNA, or protein sequences.

1.6.1. Sequence simulations

The goal of any sequence simulation is to generate a set of artificial genomic sequences that mimic, as much as possible, the characteristics of biological sequences generated by nature. In fact, what is being simulated is the process of evolution, with the aim to realistically portray the evolutionary processes that change a biological sequence (through substitutions, insertions and deletions, and more) and the functional constraints that restrict such changes. Achieving accurate simulations requires adequate theoretical models to describe how sequences are created and changed through evolution. During the simulation, the simulator keeps track of all evolutionary events and therefore can correctly determine the homology relationships for each sequence position and the "true" multiple sequence alignment.

Knowing the true evolutionary history of the sequences is the primary advantage of simulated data. It enables us to test the accuracy and robustness of many methods. Sequence simulators are often used to evaluate the performance of alignment software (Katoh and Standley, 2016), and thus play a crucial role in the quest for more accurate MSAs (Nuin, Wang and Tillier, 2006). Simulated data are also used to compare the accuracy and efficiency of phylogenetic reconstruction (Kuhner and Felsenstein, 1994; Gaut and Lewis, 1995; Huelsenbeck, 1995) and ancestral sequence reconstruction (Blanchette *et al.*, 2004). For example, to evaluate the accuracy of a method to infer the underlying parameters, we can simulate MSAs under known parameters, examining how well the method succeeded to estimate the true parameters and calculating their confidence intervals. The same can be done to estimate the null distribution for hypothesis testing (Goldman, 1993). Simulations can also be used to examine the robustness of an analytical method to model misspecification, by simulating data under a complex model and analyzing them under a simplistic incorrect model.

1.6.2. Sequences simulators

Over the years, many sequence simulators have been developed (Cartwright, 2005; Fletcher and Yang, 2009; Strope et al., 2009; Sipos et al., 2011). Here, I describe a general scheme of the simulation procedure (Fig 5) and then exemplify it in more detail according to how it is implemented in the INDELible program (Fletcher and Yang, 2009). The starting point for most sequence simulators is a phylogenetic tree with assigned branch lengths, along which the artificial evolutionary process takes place. The simulation starts at the root of the tree by

generating a root sequence according to some predefined background distribution. The evolutionary process then proceeds along each of the branches, such that evolutionary events (usually substitutions, insertions, or deletions) are drawn and applied to the simulated sequence. The longer the branch, the larger the number of events, or amount of change, expected along it. The process is stopped when all tips of the phylogeny have been reached. The result of a sequence simulation is a set of sequences for the tips of the tree as well as their true MSA. All sequence simulators are based on a specified evolutionary model that includes several aspects of evolutionary transitions. The evolutionary model is usually divided into a substitution model and an indel model.

INDELible is the most frequently used simulator for phylogenetic studies. The evolutionary process in INDELible is implemented by using the standard algorithm of simulating Markov chains, which is also known as Gillespie's algorithm (Gillespie, 1977). Here the Markov chain describes the change from one state to another where each state is the entire sequence. First, a set of exponentially distributed waiting times is specified to simulate the sequences, with mean waiting time for an event $1/\lambda$, where $\lambda = I + D + S$, such that I, D and S are the insertion, deletion and substitutions rates, respectively. At each step of the algorithm, following a branch of length T from the given tree, a waiting time t_i is sampled from the waiting time distribution. If $t_i > T$, no event occurs before the end of the branch. Otherwise, an event occurs at time t_i, and is randomly drawn to be an insertion, deletion, or substitution with probabilities $\frac{I}{\lambda}, \frac{D}{\lambda}$, or $\frac{S}{\lambda}$, respectively. The location of the event along the sequence is similarly determined by a random sampling of positions with probabilities proportional to the rates. If the event is an indel, then its length is sampled from a specified distribution. If the event is a substitution, the replacement of an element (nucleotide, codon or amino acid) is determined according to the input substitution matrix. The procedure continues to simulate events along the branch until the sum of the waiting times exceeds the length of the branch ($t_1 + t_2 + \cdots + t_n > T$).

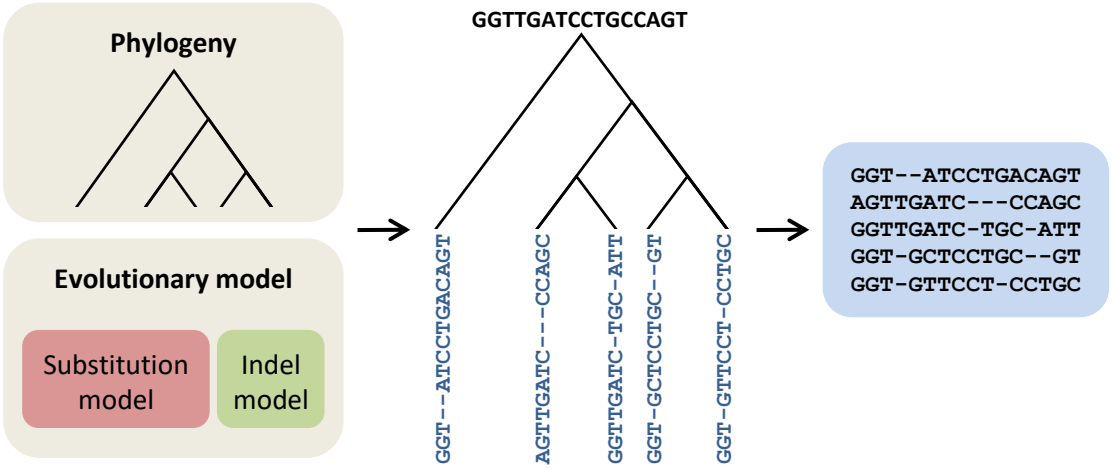


Fig 5. Schema of typical sequence simulator. The basic input is a phylogeny and an evolutionary model that usually consists of a substitution model (such as jukes-Cantor or GTR) and an indel model. Starting from a random root sequence, simulation iterates over each branch of the guide tree and applies substitution and indel events. The result is an alignment reflecting the true evolutionary path of the sequences.

1.7. Modeling indels

The nature of the simulations depends on the substitution and the indels models. While considerable research has been devoted to substitution models, such as the GTR (Tavaré, 1986) and the WAG (Whelan and Goldman, 2001) models for nucleotides and amino acids, respectively, much less attention has been paid to the development of realistic indel models. There is neither a well-established model nor a consensus regarding the distribution of indel that corresponds to a certain evolutionary distance.

1.7.1. Indel model in INDELible

In INDELilble, three parameters control the indel dynamics. The first parameter is the rate of an indel event, relative to that of substitution events. Indel can be treated as two separate events, i.e., one probability for an insertion and another for deletion, or as a single event with one probability. The second parameter is the length of the root sequence. The character composition of the root sequence depends on the type of simulation (nucleotide, amino-acid or codon) and on the base character frequencies, according to the generative model definitions (e.g., in JC69 all frequencies are fixed at 0.25). The third parameter is the distribution of the indel lengths. This distribution determines, given an indel event, the number of characters deleted or inserted. INDELible offers several possible distributions:

Zipfian, negative-binomial, and Lavalette distributions. It also has an option to specify a fixed proportion for each indel length (user-defined distribution).

1.7.2. Indels lengths distribution and the power-law distribution

Empirical studies have indicated that indel lengths follow the discrete power-law distribution (Benner, Cohen and Gonnet, 1993; Gu and Li, 1995; Zhang and Gerstein, 2003; Chang and Benner, 2004; Yamane, Yano and Kawahara, 2006; Cartwright, 2008). This distribution has a very large right tail, meaning that small indels are extremely common, whereas large indels are extremely rare (Fig 6A). Mathematically, a quantity x (here, an indel length) obeys a power-law if it is drawn from a probability distribution

$$p(x) \propto x^{-a}$$

The shape of the power-law distribution is controlled by a parameter denoted by 'a', which is in the range $(1, \infty)$. This parameter is known as the exponent or scaling parameter. The 'a' parameter is inversely related to the indel length – the lower a is, the higher the probability to get a longer indel (Fig 6B). The pure power-law distribution, known as the Zipfian distribution, is expressed as:

$$p(x) = \frac{x^{-a}}{\sum_{n=1}^{\infty} n^{-a}}$$

where the denominator is a normalizing constant, known as the zeta function.

Unfortunately, it is not straightforward to conclude with certainty whether a particular data set follows a power-law distribution due to large fluctuations that occur in the tail. Even when data are drawn from a power-law distribution, their observed distribution is extremely unlikely to exactly follow the power-law form; there will always be some small deviations because of the random nature of the sampling process. The challenge is to distinguish deviations of this type from those that arise because the data are not drawn from a power-law distribution.

Importantly, previous empirical studies that claimed that indel lengths follow the power-law distribution have not attempted to test the power-law hypothesis quantitatively. Instead, they typically relied on qualitative examination of the data, based, for instance, on visualizations (Benner, Cohen and Gonnet, 1993; Zhang and Gerstein, 2003). But such examination can be

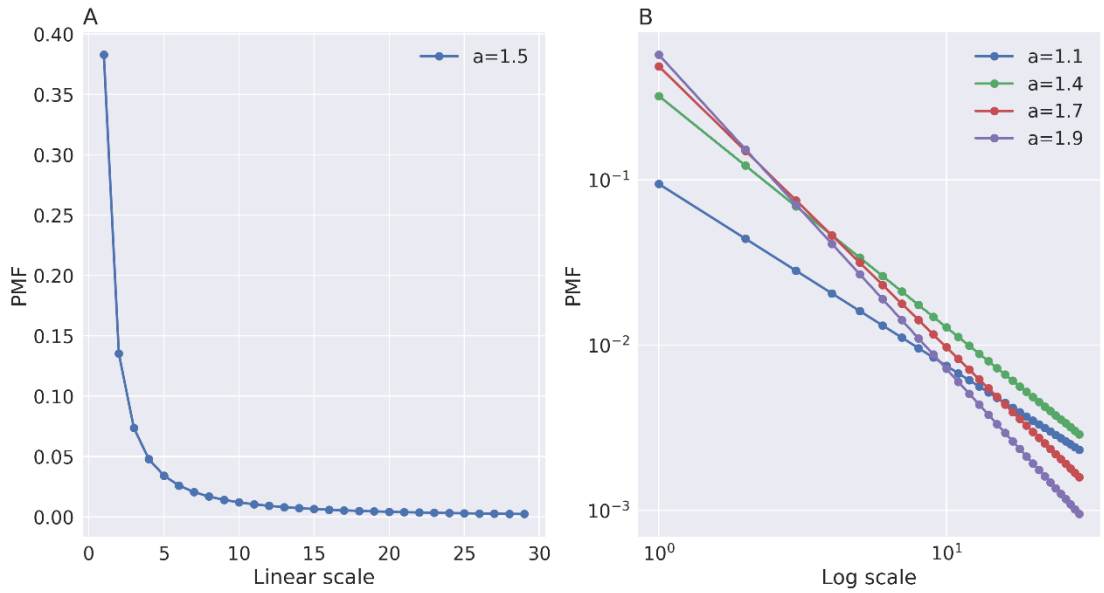


Fig 6. Zipfian probability mass function (PMF). The y axis shows the Zipfian probabilities of the x axis values. **(A)** Zipfian PMF for N = 30. **(B)** Zipfian PMF for N = 30 on a log–log scale for different values of the shape parameter 'a'. Note that the function is only defined at integer values of x. The connecting lines do not indicate continuity.

deceptive and may lead to claims of power-law behavior that do not hold up under closer scrutiny. For example, because the CDFs (cumulative distribution functions) of the normal and the exponential distributions look roughly straight in a log-log plot, one might, upon cursory inspection, judge them to follow power-law, albeit with different scaling parameters. This judgment would, however, be wrong: being roughly straight on a log-log plot is a necessary but not sufficient condition for power-law behavior. Drawing conclusions about how well the data can be approximated by a power-law distribution thus requires a quantitative measure of goodness of fit. Such a procedure was suggested by Clauset et al. (2009), and will be explained more thoroughly in the Results section.

1.8. More complex simulations

Simple models for simulations assume that the evolutionary pattern is homogenous in time and space, that is, the simulation parameters are the same along all branches or sequence sites. This is a simplistic assumption, however, and it is much more likely that some parts of the phylogeny or sequence sites evolve under different patterns compared to others. INDELible offers the ability to simulate data in multiple partitions where different partitions

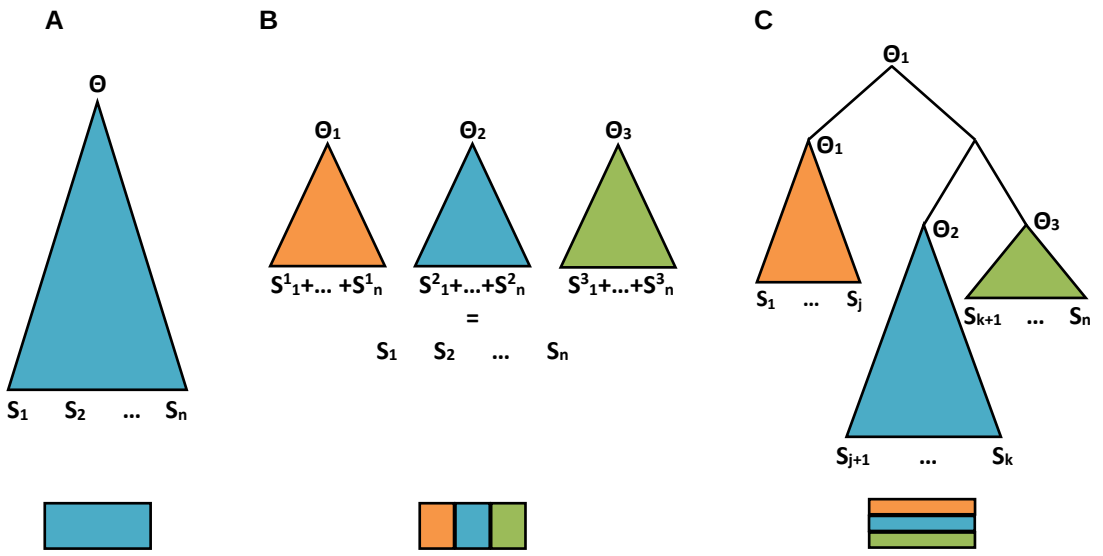


Fig 7. Scheme of complex simulations. (A) The basic simulation procedure. Given a root sequence and a global set of substitution and indel parameters (Θ), simulation proceeds by applying changes in a Gillespie manner over all sequence positions, following the guide tree and producing a set of sequences S_1 … S_n. **(B)** Simulation with multiple partitions which may have different substitution and indel parameters (Θ_1, Θ_2, Θ_3) and may evolve along different trees ($S^m{}_i$ is sequence i resulted from tree m). **(C)** A time heterogeneous simulation, where the evolution along each branch may follow a distinct model (Θ_1, Θ_2, and Θ_3). Each triangle represents a single tree structure. The bottom rectangles illustrate the resulted MSA.

may have different substitutions models, indel parameters, or may even evolve under different trees (Fig 7B). It also allows different branches of the phylogeny to follow different models (Fig 7C). Unfortunately, a main problem of the such complex possibilities is the difficulty in evaluating the parameters of the underlying models, a non-trivial task even for a simple, homogeneous model, as will be discussed next.

1.9. Parameters inference for simulations

Models allow us to simulate data, given specified values for the assumed parameters. In many biological studies, we already have the data and we want to infer the parameters that generated the data under the assumed model. For example, we want to find the parameters describing the indel dynamics, such as the ratio between indel rates and substitutions rate, and the distribution of indel lengths. Yet, estimating the indel parameters is highly challenging since this model violates the assumption of independence among sequence sites that allow efficient likelihood computations. Three notable efforts to infer indel parameters include the

"Lambda" script implemented as part of the Dawg package (Cartwright, 2005), SPARTA (Levy Karin *et al.*, 2015) and its consecutive SpartaABC (Levy Karin *et al.*, 2017).

Lambda computes several statistics from a given input MSA and estimates the indel parameters from these statistics. SPARTA searches for the set of evolutionary parameters describing indel dynamics which best fits a given input MSA. In each step of the search, it uses parametric bootstrapping to estimate how well a proposed set of parameters fits input data. SpartaABC implements an approximate-Bayesian-computation rejection algorithm to infer indel parameters from sequence data. It does so by extracting summary statistics from the data. It then performs a large number of sequence simulations under randomly sampled indel parameters. By computing a distance between the summary statistics extracted from the empirical data and the simulated ones, SpartaABC retains only parameter values for simulations that are close to the real data, and then compute some statistic on these (e.g., the mean). Using simulated data, SpartaABC was shown to outperform the other two methods.

1.10. Simulation realism

As stated before, the goal of sequence simulation is to generate sequences that resemble as much as possible the real data. One way to test how much this aim has been achieved is to use tools of alignment assessment. Originally, these tools were developed with the aim to measure alignment reliability; however here they are used as a measure of alignment complexity, with the rational that the real and the simulated data should have the same level of complexity. One such tool is GUIDANCE2 (Sela *et al.*, 2015). This method assigns a confidence score for each column in an input alignment (the higher the score, the higher the confidence in the alignment of a specific column), which can be used for weighting or filtering unreliability columns in the alignment. These scores can be interpreted as a measure of alignment difficulty: the lower the confidence in the alignment of a specific column, the more complex its computation.

Below I provide a brief description of the GUIDANCE2 procedure (Fig 8). First, a standard MSA of the input sequences is computed, hereby termed "base MSA". Second, a set of MSAs is constructed. Each of these MSAs is generated based on an alternative guide-tree (produced through non-parametric bootstrap) that is provided to the alignment algorithm, as well as

alternative gap penalties. Third, the set of alternative MSAs are compared to the base MSA in order to estimate its confidence level, resulting in scores between 0-1 for each column. A column score indicates the percentage of times the column appeared in the alternative MSAs (i.e., how many times the column is aligned the same in the alternative MSAs, divided by the number of alternative MSAs).

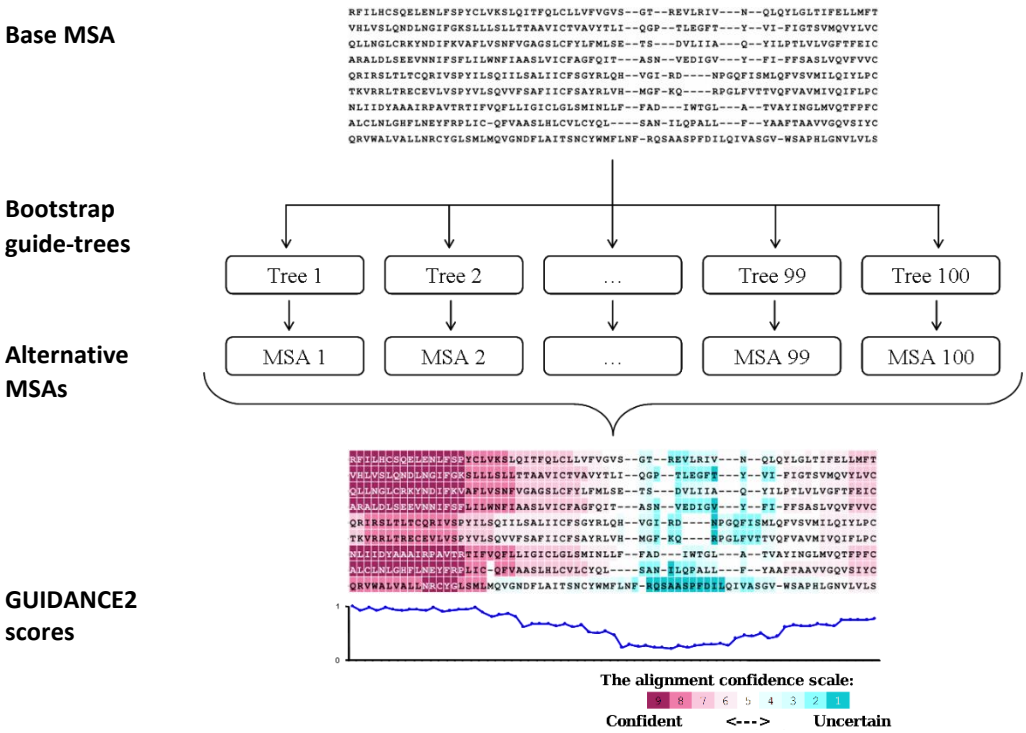


Fig 8. A schematic flowchart of the GUIDANCE algorithm. A base MSA is first produced by an alignment program specified by the user. Bootstrap trees are reconstructed and given as guide trees to the alignment program, producing a set of MSAs. GUIDANCE2 scores are then calculated by comparing each MSA to the base MSA, and are color coded on each residue in the alignment according to the color scale at the bottom right.

2. Research objectives

In this study I aimed to evaluate several components that are central to the reconstruction of mega-phylogenies:

- Examining the accuracy of large published phylogenies, using the plant mega phylogeny as a test study.
- Testing the effect of alignment complexity on the accuracy of inferred phylogenies using sequence simulations approach.
- Evaluating the realism of simulated alignments by comparing their complexity to that of empirical data.
- Generating realistic alignments that resemble the indel characteristics of the biological data.

3. Results

3.1. Assessing monophyly in a published large phylogeny

I first aimed to evaluate the accuracy of mega phylogeny inference, as reconstructed using the super-matrix approach. To this end, I examined several large phylogenies and measured their agreement with accepted taxonomical classification. Specially, I tested whether species that are circumscribed under several taxonomic hierarchies (e.g. genus, family, Fig 9) appeared as monophyletic groups in the tree, and, in case they are not monophyletic, what the extent of disagreement is. The test for monophyly involved two parts: first, for a given taxonomic rank, all the tips contained within that rank were identified in the tree (usually, not all taxa defined under a certain taxonomic classification are present in the tree). Second, the tree topology was checked to determine if the set of tips formed a monophyletic group, that is, they were all contained within a single subtree, and no other tips are contained within this subtree. If this condition was met, I inferred that the taxonomical classification for the focal group perfectly agrees with the phylogeny. Conversely, if this condition was not met, two relaxed measures were used to describe the level of disagreement (Fig 10). The first score is the number of species in the focal taxonomic group, divided by the number of species that appear in the phylogeny in the sub-tree that includes the most recent common ancestor (MRCA) as its root. The second score is the number of species in the largest monophyletic subgroup (i.e., the largest group that only contains species from the focal group), divided by the number of species in the group. For both scores, values approaching 1 are indicative of high concordance between the taxonomic classification and the inferred phylogeny, while scores approaching zero are indicative of low overlap. The need for two different scores is exemplified in Fig 11.

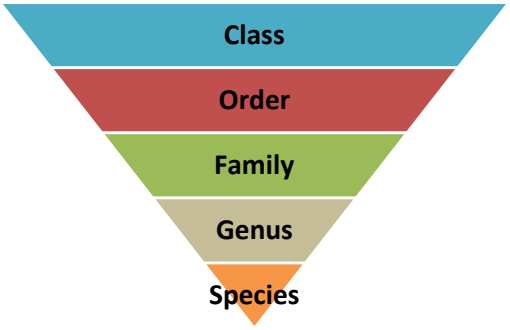


Fig 9. Examples of taxonomic ranks.

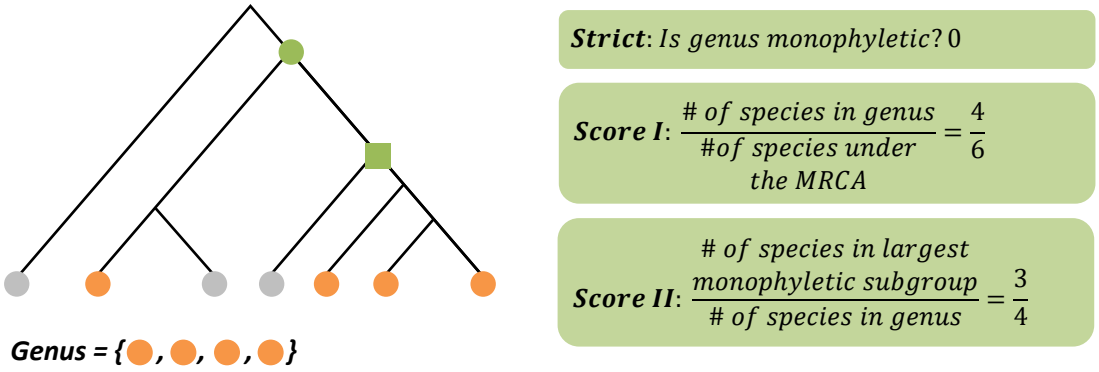


Fig 10. Example of calculating monophyly scores given a tree and a focal genus. Species are represented by circles. The focal genus contains 4 species (orange circles). The MRCA of the focal genus is marked by a green circle. In this case, the focal genus does not appear as a monophyletic group in the inferred tree since two species that belong to another genus subtend the MRCA. The corresponding monophyly scores are 0.66 and 0.75, respectively. The first score is obtained by dividing the size of the focal genus (4) by the number of species under the MRCA (6). The second score is obtained by dividing the largest monophyletic group size (3) (which its node is indicated by a green square) by the genus size (4).

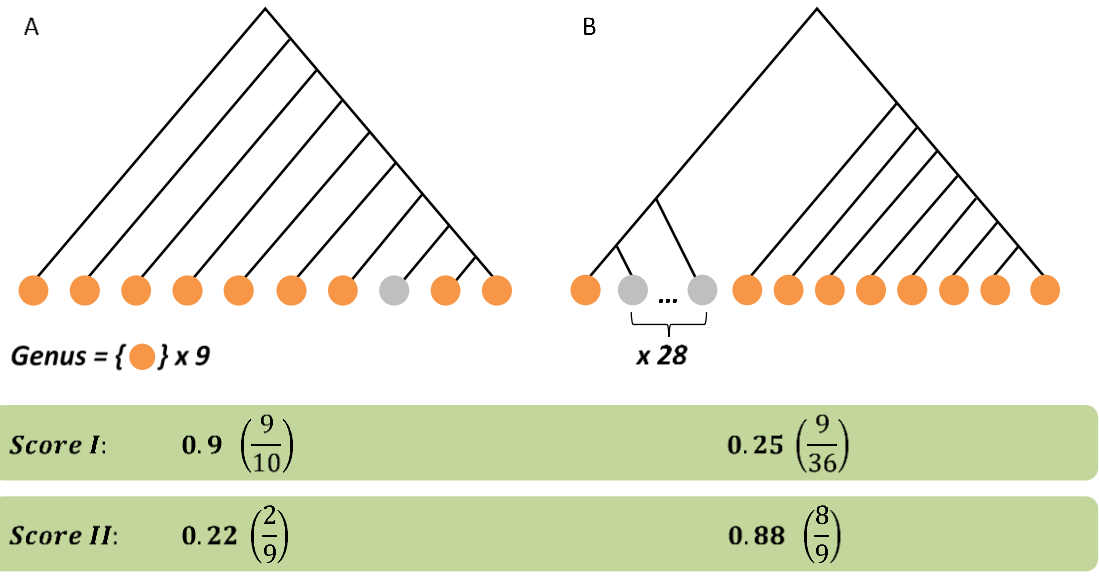


Fig 11. Two illustrative cases where the two monophyly scores greatly differ. A genus consisting of nine species is shown in two phylogenies. **(A)** The MRCA contains only one species that does not belong to the genus, resulting in a high value of the first score, while the smallest monophyletic group is of size two, resulting in a low value of the second score. **(B)** The last common ancestor contains 28 species that do not belong to the genus, resulting in a low value of the first score, while the largest monophyletic group includes all but one species of the genus, resulting in a high value of the second score.

To evaluate the extent of discrepancy between inferred phylogenies and current taxonomical knowledge, I used the large Magnoliophyta (angiosperms) clade, which includes roughly 300,000 species (Christenhusz and Byng, 2016). A large angiosperm phylogeny was reconstructed by Zanne et al. (2014), encompassing ~32,000 land-plants species, out of which 28,224 belong to the Magnoliophyta. Using this tree, I computed the two monophyletic scores for each recognized group at three levels of taxonomical hierarchies: genera, families, and orders. For each hierarchy, I computed the average over all clades examined (e.g., the results at the genus level were combined from all 528 genera with more than 10 taxa). Notably, the phylogeny reconstructed by Zanne et al. included topological constrains on all families and orders (but not genera), meaning that the groups of these ranks are expected to be monophyletic or close to monophyletic in the inferred phylogeny. Therefore, I repeated the reconstruction of the phylogeny with the same set of 28,224 species and the same genetic markers, using two alternative procedures. First, I used the PHLAWD pipeline, which accounts for the taxonomy when aligning the sequences (see Introduction) and applied ExaML on the resulting alignment to reconstruct the tree. Second, I realigned the sequences with MAFFT and applied ExaML for phylogeny reconstruction. The strict and the two relaxed monophyletic scores were then computed for each of the three mega phylogenies (termed Zanne, PHLAWD, and MAFFT). Notably, these three phylogenies account for the taxonomical relationships to different extent: Zanne's tree includes taxonomical constraints as part of the tree reconstruction process, PHLAWD as part of the alignment, and MAFFT does not include any taxonomical constraints (see Methods, section 5.3).

As can be seen in Fig 12, roughly 40% of all genera analyzed are strictly monophyletic. This result is consistent across all three phylogenies. Similar monophyly percentages were obtained at the family rank for the two phylogenies reconstructed here (PHLAWD and MAFFT), while a much lower percentage was obtained at the order level. In Zanne's tree, nearly all families and orders were inferred as monophyletic groups. This is expected due to the topological constraints that were employed when reconstructing the tree.

Considering the two relaxed monophyly scores, low values were obtained for score 1 for PHLAWD and MAFFT trees, particularly at the order level (Fig 12). This indicates that the MRCA of each focal group includes many species that do not belong to this group according to

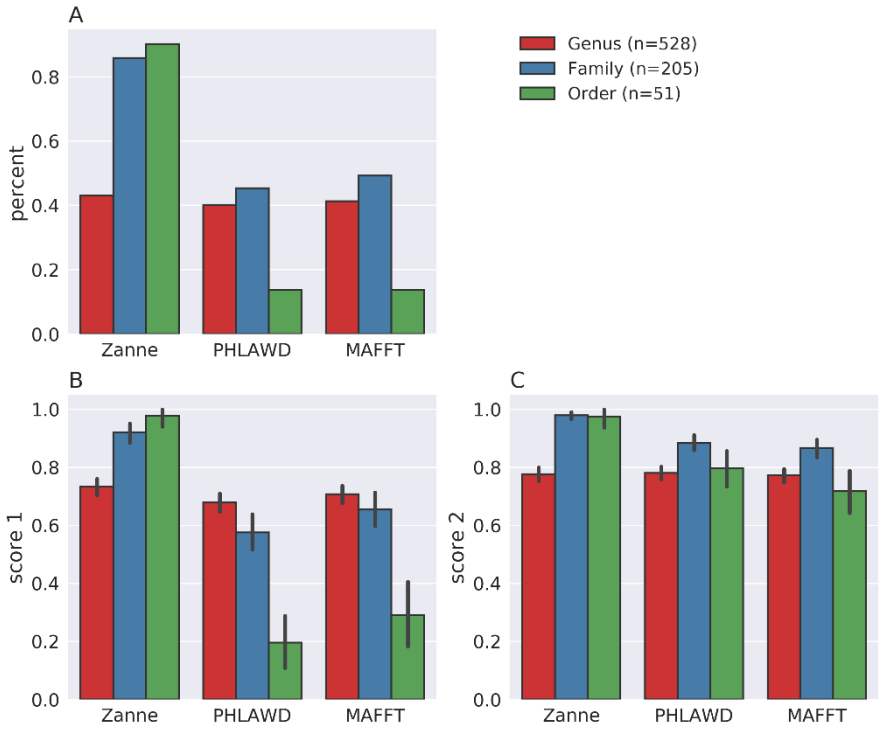


Fig 12. Monophyly scores for Magnoliophyta phylogeny reconstructed in three ways. Scores are calculated for three taxonomic ranks: order, family and genus. The size of each group is indicated in the legend. **(A)** Percentage of monophyletic (strict definition) groups within each rank. **(B)** Average of monophyly score 1. **(C)** Average of monophyly score 2. The analysis was conducted only for groups with more than 10 species.

current taxonomical knowledge. On the other hand, the average values of score 2 were above 0.8 for all trees and ranks examined. High values of score 2 indicate that a large fraction of the focal group is monophyletic. Taking together, the difference between the two relaxed monophyletic scores implies that in many cases a large fraction of the focal group is monophyletic and there are few species that appear further away in the reconstructed phylogeny, such that many additional species are included in the sub-tree that encompasses all in-group species.

One would expect that the phylogeny reconstructed based on the PHLAWD alignment would be more congruent with the taxonomy compared to the phylogeny reconstructed based on the MAFFT alignment due to the consideration of the taxonomy when computing the alignment. However, my results indicate that the two phylogenies do not differ substantially (Fig 12). A possible explanation is that PHLAWD aligned much more sequences when it

compiled and aligned the sequences of all species in Magnoliophyta, and only then species that were not found in Zanne's tree were removed, while the MAFFT alignment considered only the relevant sequences (i.e., those sequences that belong to species that are present in Zanne's tree). It is also possible that the heuristics implemented in PHALWD do not meet their aim to improve the quality of large alignments based on taxonomical knowledge. To sum up, without forcing constrains on the tree topology, less than half of the analyzed clades were found monophyletic in the inferred phylogenies. Assuming that at least some of these incongruences were due to the quality of the MSA, the next step was to examine the effect of MSA complexity on the accuracy of the inferred phylogeny.

3.2. Testing the effect of alignment complexity on the accuracy of inferred phylogenies

The analyses detailed above revealed that for large phylogenies, such as the one reconstructed for all land plants, a large fraction of recognized clades are not monophyletic. I hypothesized that the root cause for the low agreement between the reconstructed phylogenies and current taxonomical knowledge lies in current methods of MSA construction that produce erroneous alignments when sequences that are too far apart are assembled together.

To test this hypothesis, I used simulations where the accuracy of inferring the phylogeny of a focal group was examined in the context of increasingly larger alignments. Briefly, a large phylogeny served as a reference tree and a subgroup of species was chosen as the focal group. Increasingly larger sets of species belonging to the reference tree were selected and sequences were simulated along the reference tree for these species. For each simulation, phylogenies were inferred once based on the true MSA (i.e., the exact alignment that was the result of the simulation) and once based on an MSA that was re-aligned using MAFFT. The inferred and reference trees were filter to only include species belonging to the focal group, and their distances were computed. Fig 13 illustrates this process and a detailed description is provided in Methods (section 5.4). In this procedure, if errors in the alignment procedure have an effect on the accuracy of phylogeny reconstruction, then we expect that the trees inferred based on the MAFFT alignment will be further away from the reference trees compared to those inferred based on the true alignment. Additionally, if errors in the alignment are more prevalence as the extent of divergence among sequences increases, then

we expect that the resulting distances will be larger as the size of the MSA increases (note that following the filtering step, the size of the compared trees are the same in all comparisons).

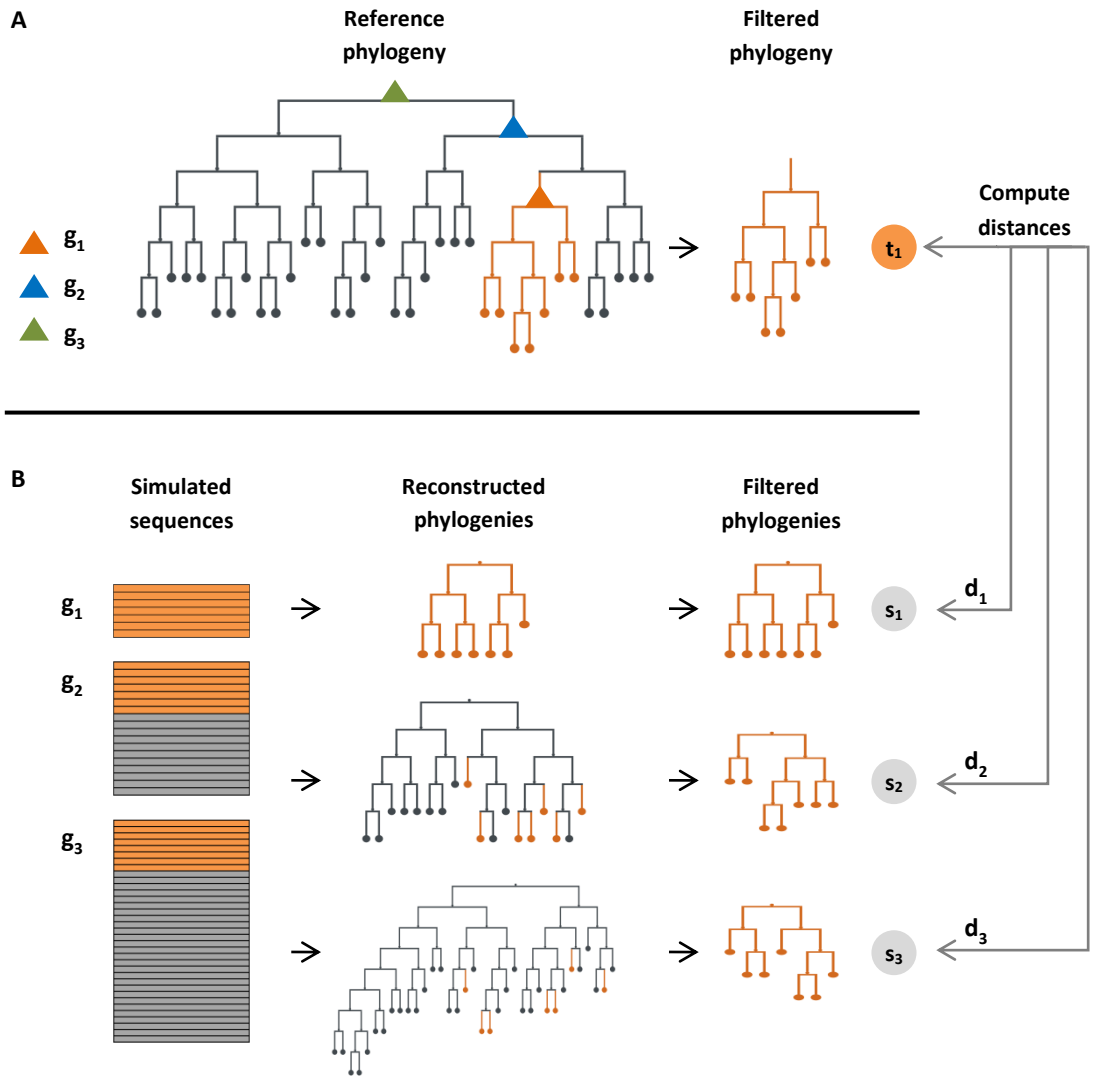


Fig 13. Examining the accuracy of phylogeny inference in the context of increasingly larger phylogenies. (A) Starting from a reference phylogeny, three monophyletic groups that are contained within each other (denoted g_1, g_2 and g_3) were chosen, and g_1 was pruned to form the reference sub-tree t_1. **(B)** Sequences were simulated along the reference tree, to generate three groups of simulated sequences, corresponding to g_1, g_2 and g_3. These sequences were used to reconstruct the corresponding phylogenies. From each of the inferred phylogenies, all species were filtered, except those belonging to g_1 to obtain the filtered sub-trees s_1, s_2, and s_3. Finally, the distance between t_1 and s_1, s_2 and s_3 were computed (denoted d_1, d_2, and d_3, respectively). The entire procedure was performed to each of the focal groups; for simplicity, the process is illustrated here for g_1 only.

Table 1. Distances between subtrees of the reference phylogeny and those reconstructed from simulated data. The first column specifies the size of reference subtree; while columns 2-5 list the mean distances between the reference subtree inferred in the context of the larger phylogenies by filterinfg species not part of the reference subtree. Distances were computed using the Branch-score distance. Each entry represents the mean tree distance over 10 independent simulations, while the standard deviation is given in parentheses. The subscript in the referenced and inferred trees (t_x and s_x, respectively) indicates the number of species in the phylogeny.

	Reference tree	Inferred tree				
		s_{135}	s_{816}	s_{2412}	s_{5113}	*p-value*[1]
True MSA	t_{135}	0.37 (0.02)	0.43 (0.02)	0.43 (0.02)	0.44 (0.02)	5.6×10^{-6}
	t_{816}		0.63 (0.02)	0.64 (0.02)	0.66 (0.02)	0.01
	t_{2412}			0.88 (0.01)	0.90 (0.01)	0.06
	t_{5113}				1.10 (0.01)	
MAFFT MSA	t_{135}	0.38 (0.02)	0.43 (0.02)	0.43 (0.03)	0.47 (0.04)	1.7×10^{-6}
	t_{816}		0.63 (0.02)	0.64 (0.02)	0.68 (0.02)	1.8×10^{-4}
	t_{2412}			0.89 (0.02)	0.92 (0.02)	1.7×10^{-3}
	t_{5113}				1.13 (0.02)	

[1]*p-value* obtained using paired-ANOVA

The results revealed a trend whereby trees inferred based on smaller alignments were more similar to the reference tree than those based on larger alignments (Table 1). For example, the first row in Table 1 shows the distances between the reference phylogeny of a focal group that included 135 species with those inferred within the context of four increasingly larger phylogenies. While the mean distance between the reference tree and the tree that was inferred based on an MSA that only included the 135 species belonging to the focal group is 0.37, the mean distances of the inferred trees that are based on MSAs of sizes 816, 2,412, and 5,113 are 0.43, 0.43, and 0.44, respectively (p-value << 0.01; paired ANOVA; the most significant difference is between the largest and smallest trees based on Tukey's test). It thus seems that the most substantial difference was between the smallest MSAs of size 135 and all the larger MSAs, which included at least six times more species. However, the results obtained using the true MSA and that computed using MAFFT were highly similar. Additionally, while the trees inferred from larger phylogenies were significantly less accurate than those inferred from smaller ones, the magnitude of the difference in the distances shown in Table 1 is smaller than I expected. Together, these findings suggest two alternative explanations. One possibility is that the major source of errors in large phylogeny reconstruction is due to inaccuracies in

tree reconstruction methods while the alignment computation has a minor effect. Alternatively, it is possible that no differences were obtained between the two types of inferred trees (based on the true and MAFFT MSAs) since the simulated MSAs that were generated using the INDELible program were not realistic enough and produced sequences that are not as challenging to align. According to this scenario, the construction of the MSAs was not affected by the number of species and by their evolutionary distances at an extent that mimic empirical data. This last possibility will be examined in the section below.

3.3. Do we simulate realistic sequences?

There are two central issues when one aims to simulate sequence data that mimic patterns observed in reference sequence data. First, we need to determine the parameters values that govern the simulation model. Second, we need to determine how to compare whether the complexity of the simulated alignments resemble that of the reference data.

3.3.1. Estimating the parameters for sequence simulations

The INDELible simulator that was used in my study requires the specification of substitution and indel models. Yet, while the parameters of the substitution model can be inferred using standard maximum likelihood procedures (see Methods, section 5.5), those of the indel model are more challenging to estimate since incorporating indel events within the likelihood function violates the assumption that different sites evolve independently, which allows efficient likelihood computations (Felsenstein, 1981). There are three indel parameters that are needed to be estimated (Fig 14): 'RL', representing the root length; 'IR', representing the indel rate relative to the substitution rate; 'a', representing the shape parameter of the power-law distribution that controls the indel-length distribution. I thus investigated several alternative procedures for this inference task.

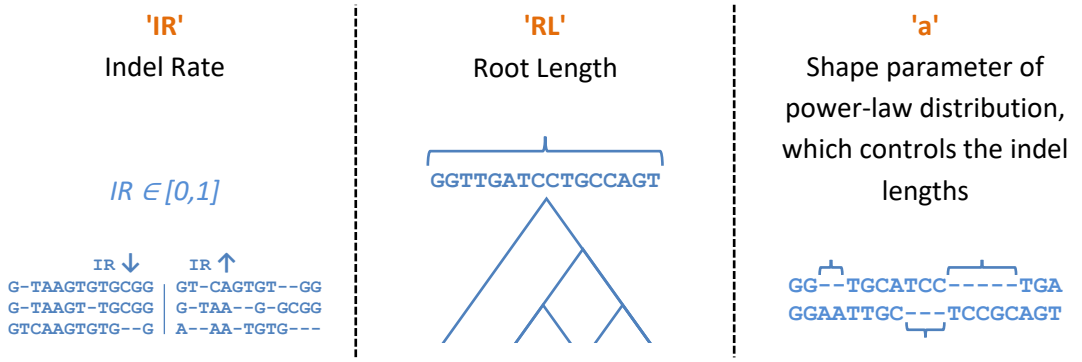


Fig 14. Illustration of the three Indel parameters used in INDELible simulator. The parameters are: indel rate ('IR') - the rate in which an insertion or deletion event occurs in the sequence; root length ('RL') - the length of the root sequence which serves as the starting point for the simulated evolutionary process; the shape parameter of the power-law distribution ('a'), which controls the lengths of the inserted or deleted sequences.

3.3.1.1. Estimating the parameters for sequence simulations: existing tools

I first compared between three available methods that estimate the three indel parameters: Lambda (Cartwright, 2005), Sparta (Levy Karin *et al.*, 2015), and SpartaABC (Levy Karin *et al.*, 2017). For each method, the indel parameters were estimated from the reference alignments for each marker separately using two different datasets of 132 and 5,113 species (termed small dataset and large dataset, respectively; see Methods, section 5.2). For each dataset, seven markers were analyzed; these markers are the one used by Zanne et al. to reconstruct the mega phylogeny of plants. As can be seen in Table 2, there are large discrepancies in the parameter values estimated by the three tools. For example, the 'IR' and 'RL' values inferred by SpartaABC were substantially smaller compared to those inferred by Lambda. This was the case also for the 'a' values in the small data set but not in the large dataset. In addition, large discrepancies were observed between SPARTA and the two other methods, particularly with regard to the 'IR' parameter. Finally, the 'IR' value inferred by SpartaABC for *atpB*, *matK*, and *rbcL*, was 0, meaning the absence of indel events, an inference which contradicts the pattern observed in the input data.

Next, I examined whether MSAs that were simulated based on the inferred parameters resemble the indel characteristics of the reference MSAs. To this end, I examined whether the reference MSAs and the simulated MSAs are similar with respect to their total length and the average number of gap blocks per sequence, where a gap block is defined as a stretch of

Table 2. Values of indel parameters inferred by three methods. Each row specifies the inferred values of the three indel parameters (IR, a, RL) by the three examined methods, when given an MSA of the specified marker. Parameters of the large data set could not be inferred by SPARTA due to the size of the input data matrix.

		Lambda			SPARTA			SpartaABC		
	Marker	**IR**	**a**	**RL**	**IR**	**a**	**RL**	**IR**	**a**	**RL**
Large data set	***18S***	0.002	1.353	1,654	-	-	-	0.001	1.838	226
	26S	0.004	1.336	1,997	-	-	-	0.001	1.739	571
	ITS	0.051	1.354	617	-	-	-	0.002	1.387	275
	atpB	0.001	1.222	1,412	-	-	-	0	1.446	779
	matK	0.005	1.288	1,540	-	-	-	0	1.491	1,117
	rbcL	0.002	1.214	1,295	-	-	-	0	1.42	918
	trnL-F	0.028	1.341	820	-	-	-	0.002	1.104	440
Small data set	***18S***	0.005	1.466	1,689	0.061	1.566	1,805	0.001	1.067	937
	26S	0.011	1.428	2,350	0.137	1.265	3,365	0.001	1.056	1,346
	ITS	0.048	1.298	633	0.083	1.464	1,155	0.003	1.043	492
	atpB	0.003	1.26	1,440	0.146	1.55	1,456	0.001	1.068	691
	matK	0.007	1.335	1,529	0.145	1.539	1,791	0.001	1.064	1,369
	rbcL	0.002	1.287	1,381	0.094	1.55	1,427	0.001	1.069	869
	trnL-F	0.049	1.377	836	0.019	1.508	633	0.005	1.065	631

consecutive gap characters. Evidently, the simulated MSAs highly deviated from the reference alignments with respect to their length and number of gaps (Table 3). The high parameter values inferred by Lambda led to longer MSAs with higher number of gaps compared to the reference MSAs. For example, the lengths of the simulated *ITS* alignments for the small and large datasets were 5,463 and 55,075, respectively, roughly 5- and 20-fold longer than the reference alignments. The indel values inferred by SpartaABC, on the other hand, lead to substantially shorter MSAs compared to the reference MSAs, although the number of gap blocks was generally higher. These results thus indicate that the use of existing methods for indel-parameters estimation fail to generate MSAs that resemble reference MSAs in their indel characteristics, at least for the set of sequence data sets used in this study. I thus aimed to develop my own procedure for this task.

Table 3. Length and average number of gaps of reference and simulated MSAs. MSAs were simulated by INDELible with indel model parameters inferred by three methods: Lambda, Sparta, and SpartaABC. Statistics are an average over 50 simulated MSAs. Len is the alignment length and Gaps is the average number of gap blocks in the MSA. The statistics obtained using Sparta for the large data set are missing since this method failed to run on the given MSA.

		Reference MSA		Simulated MSAs					
				Lambda		Sparta		SpartaABC	
	Marker	Len	Gaps	Len	Gaps	Len	Gaps	Len	Gaps
Large data set	***18S***	1,836	26	4,795	399	-	-	344	33
	26S	3,715	60	7,595	645	-	-	848	68
	ITS	2,205	87	55,075	592	-	-	1,154	107
	atpB	1,521	4	3,506	227	-	-	779	0
	matK	2,164	29	10,623	806	-	-	1,116	0
	rbcL	1,433	5	5,713	432	-	-	917	0
	trnL-F	1,938	104	33,623	753	-	-	1,748	117
Small data set	***18S***	1,813	10	2,694	162	13,854	1,163	1,176	21
	26S	3,494	31	5,927	495	81,962	2,227	1,776	40
	ITS	1,041	26	5,463	341	13,761	806	846	16
	atpB	1,517	3	2,135	87	25,549	1,110	1,064	19
	matK	1,891	16	3,053	198	31,752	1,357	1,658	30
	rbcL	1,427	3	1,876	65	16,444	1,029	862	33
	trnL-F	1,326	55	6,807	476	2,083	204	1,352	64

3.3.1.2. An alternative procedure to estimate indel parameters

I developed an optimization-based approach, termed OPTIMIM (Optimize Indel Model), for the inference of the indel parameters given a multiple sequence alignment. A schematic workflow of the OPTIMIM methodology is provided in Fig 15. This method searches over the space of indel parameters and chooses the set that produces MSAs that most resemble the input MSA. Similarity between two alignments is defined here with respect to the number of gap blocks and the total MSA length – the two parameters that describe most succinctly an examined MSA. The procedure first set the root length parameter ('RL') to be the average sequence length of the input sequences. It then iterates over a large number of possible values for the 'IR' and 'a' parameters. This is done by setting a grid of possible combinations. In all analyses described here, the range of possible values were set as $[10^{-6}, 0.01]$ and (1.0, 1.6], for the 'IR' and 'a' parameters, respectively, each divided to 30 intervals evenly spaced on a log scale. For each parameter set, 30 independent simulations were performed. The fit of each

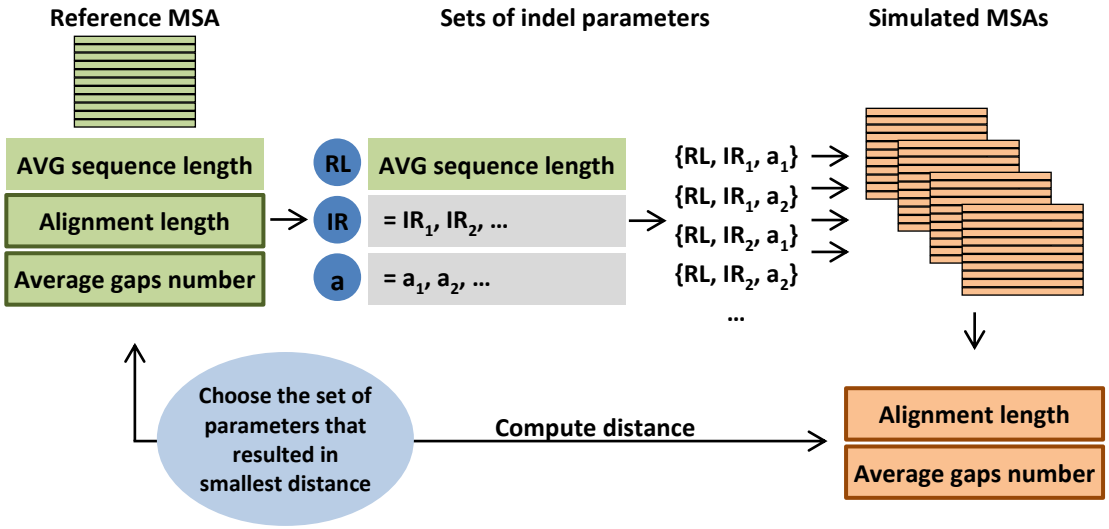


Fig 15. Scheme of OPTIMIM method to infer indel parameters. Starting from a reference MSA, three statistics are extracted: average sequence length, total alignment length, and average number of gap blocks. Following, sets of indel parameters are defined, where the RL is set to the average length of the reference sequences, and 'IR' and 'a' are combinations of pre-defined values. For each set of parameters, a corresponding MSA is simulated under this set and the alignment length and the average number of gaps are extracted from each alignment. The distance of these statistics from those of the reference MSA are computed, and the set of parameters which led to the smallest distance is the inferred one.

set of parameters to the input MSA was computed as the average distance between the input MSA and the simulated MSAs produced under this set of parameters. The distance, denoted by $d(Ref, Sim)$, was defined as the sum of the relative errors between the two summary statistics (MSA length and average number of gaps), when comparing the simulated and reference MSAs:

$$d(Ref, Sim) = \frac{\left| Len_{Ref} - Len_{Sim} \right|}{Len_{Ref}} + \frac{\left| GAP_{Ref} - GAP_{Sim} \right|}{GAP_{Ref}}$$

where Len_{Ref} and Len_{Sim} are the length of the reference and simulated MSAs, respectively, and, similarly, GAP_{Ref} and Gap_{Sim} are their average number of gap blocks.

I next used the OPTIMIM method to infer the indel parameters for each of the reference alignments (Table 4). There are two important differences between the inferences obtained using OPTIMIM and the other existing methods (compare Table 4 to Table 2). First, the

Table 4. Inference of indel parameters using the OPTIMIM method, and length and average number of gaps of the simulated MSAs using these parameters. Statistics are an average over 50 simulated MSAs.

	Marker	Inferred parameters			Reference MSA		Simulated MSAs	
		IR	a	RL	Len	Gaps	Len	Gaps
Large data set	***18S***	0.00012	1.00000041	1,659	1,836	26	1,986	30
	26S	0.00030	1.00000012	2,002	3,715	60	2,744	65
	ITS	0.00079	1.00000259	618	2,205	87	1,850	102
	atpB	0.00001	1.00000001	1,415	1,521	4	1,450	3
	matK	0.00012	1.00000481	1,542	2,164	29	1,866	30
	rbcL	0.00057	1.00003067	821	1,433	5	1,345	4
	trnL-F	0.00012	1.00000041	1,659	1,938	104	1,877	89
Small data set	***18S***	0.0002	1.00000002	1,698	1,813	10	1,794	9
	26S	0.0006	1.00000022	2,358	3,494	31	2,668	29
	ITS	0.0020	1.00000022	634	1,041	26	951	28
	atpB	0.0001	1.00000075	1,448	1,517	3	1,471	2
	matK	0.0006	1.00067269	1,530	1,891	16	1,733	18
	rbcL	0.0001	1.00001654	1,383	1,427	3	1,410	3
	trnL-F	0.0028	1.05073443	837	1,326	55	1,409	50

inferred parameters were markedly different. Specifically, the 'RL' parameter inferred by OPTIMIM closely resembled that inferred by Lambda and was generally higher than that inferred by SpartaABC. The 'IR' and 'a' parameters were generally smaller than those inferred by the other methods. As an exception to this pattern, for three markers (atpB, matK, and rbcL in the large dataset) for which SpartaABC inferred the 'IR' parameter to be unrealistically zero (see above), the 'IR' parameter inferred by OPTIMIM was higher. Second, MSAs that were simulated based on the indel parameters inferred by OPTIMIM resembled more closely the reference MSAs in terms of total length and average number of gaps compared to MSAs that were generated based on parameters inferred with previous methods (compare Table 4 to Table 3).

3.3.2. Evaluating the complexity of the simulated alignments

Once I could generate simulated MSAs that resemble these of the reference alignments using the OPTIMIM method, I examined the hypothesis that simulated data are much easier to align compared to real biological sequences. To this end, I used alignment reliability scores as computed using GUIDANCE2. The method assigns a confidence score in the range of (0,1) for each column in an alignment (the higher the score, the higher the confidence in the alignment of a specific column) and these can be interpreted as a measure of alignment difficulty (see Introduction, section 1.10). A GUIDANCE2 score for the entire alignment is the average of the GUIDANCE2 scores over all columns in it. This comparison demonstrated that in nearly all cases examined, GUIDANCE2 scores of the simulated alignments are substantially higher than those obtained for the reference sequences (Fig 16). The method by which the indel parameters were inferred did not have a consistent effect on the alignment complexity (i.e., for some markers, GUIDANCE2 scores were higher using parameters inferred using OPTIMIM compared to those inferred using SpartaABC and for other markers the opposite was observed). These results thus demonstrate that in the vast majority of the cases, simulated sequences do not resemble the complexity observed in real biological data and thus cannot be used to generate complex alignments that are needed to evaluate the analyses of evolutionary diverged sequences.

3.3.3. Increasing simulation complexity

As seen in the previous section, sequences obtained using simulations are not as complex as those of real biological data, at least as indicated by the GUIDANCE2 measure. Thus, I examined three possible ways to solve this issue.

3.3.3.1. Examining simulated sequence lengths

First, I examined the distribution of the sequence lengths, created by the process of insertions and deletions. Comparing the length distribution of the simulated sequences against that of the reference sequences revealed that the variability among the simulated sequences is substantially lower than that of the reference sequences (Fig 17). For example, the coefficient of variation (CV) of the reference sequence lengths of the *atpB* marker is 0.09, while that of the simulated sequences is 0.001 (average value over 50 simulations). For the *26S* marker, the

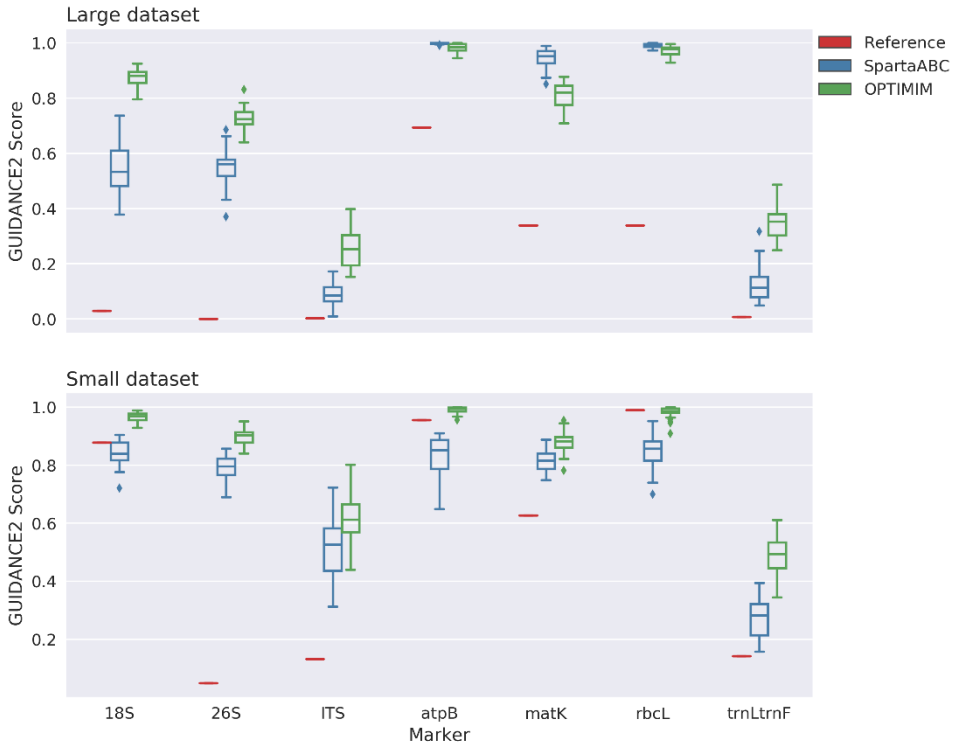


Fig 16. Alignment complexity of real and simulated data. For each marker, two boxplots of GUIDANCE2 scores are presented, differ in the way the input indel parameters were inferred, whether by using the SpartaABC (blue bars) or by OPTIMIM (green). The red horizontal bar to the left of each pair of box plots represents the score obtained for the reference MSA. Because of the limit of GUIDANCE2 on the size of the input sequences, the number of sequences of the large dataset was limited to 1200. The sequences were chosen in a manner that would maximize their sequence divergence using Phylogenetic Diversity Analyzer software (Swenson, 2009). Scores are obtained for 30 simulated MSAs.

phenomenon is even more extreme - the CV of the sequence lengths of the reference sequences is 0.55 while that of the simulated sequences is 0.006.

The difference in sequence lengths variability between the real and simulated MSAs could be explained by the oversimplified model that was used to simulate the data and by the fact that my analysis did not account for non-evolutionary factors. Namely, many of the sequences that are available in GenBank were not sequenced to their whole length but represent only partial sequences. For example, many of the sequences of the *26S* marker were only partially sequenced (as can be seen in the sequences description of the retrieved sequences). The inclusion of short sequences resulted in long indel at both edges of the alignment, and this may affect the accuracy of the alignment.

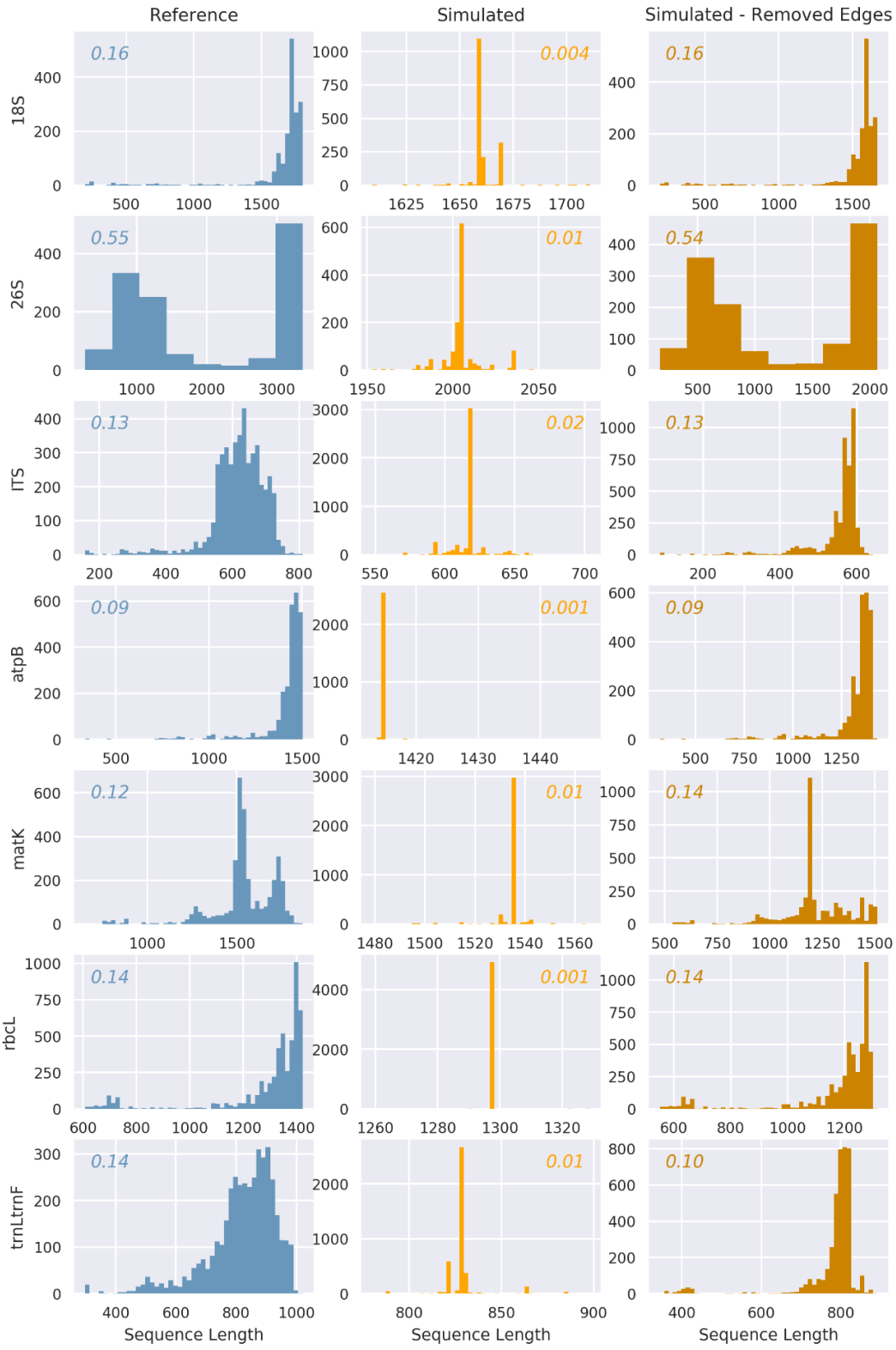


Fig 17. Histograms of sequence lengths for reference and simulated alignments of large dataset. Alignments were simulated with indel parameters inferred by OPTIMIM. The simulation whose histogram is shown here is the one with median SD over out of 50 simulations. Italic numbers at the top corner of each panel represent the coefficient of variation. The histograms of the small dataset exhibit a similar pattern (not shown here).

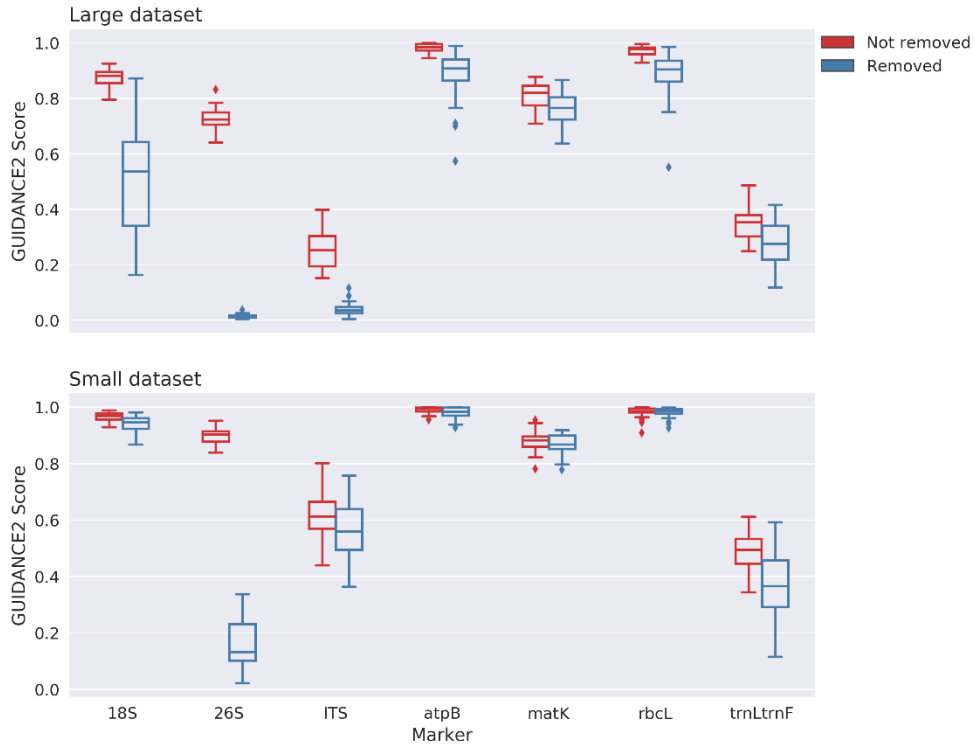


Fig 18. Alignment complexity of simulated data following the removal of edges. The edges of the simulated data (simulated under the indel parameters inferred by OPTIMIM) were removed, in the same proportion as in reference data. For each marker, a boxplot of GUIDANCE2 scores of the resulted sequences is presented (blue bars). The boxplot is compared to the boxplot of simulated data in which the edges were not removed (red bars). For each marker, scores were obtained for 30 simulated MSAs. The size of the input sequences of the large dataset was limited to 1200.

To test the effect of partial sequences on alignment complexity, I mitigated the effect of partial sequences on the simulated MSAs. To this end, I removed nucleotide characters from the edges of the simulated sequences in the same proportion as in the reference MSA. For example, if the first sequence in the reference MSA had gaps of length 20% and 30% of the MSA length at the beginning and end edges, respectively, then the corresponding simulated sequence was clipped at the beginning and the end in the same proportion. GUIDANCE2 score was then computed for the clipped sequences. Evidently, this procedure resulted in lower GUIDANCE2 scores for all markers and datasets analyzed (Fig 18). Moreover, the decrease was particularly noticeable for *26S*, the marker with highest sequence length variability, indicating that the presence of a large number of partial sequences leads to alignments that are more difficult to align.

3.3.3.2. Non-homogeneous indel model

Another possible explanation for the over simplified MSA produced via simulations could be that the indel process is not homogeneous across the alignment. Namely, while in reality some parts may be indel rich (and thus hard to align) and some are indel poor, the simulation model assumed that indels are spread homogenously across the alignment. To examine whether more complex alignments could be obtained via heterogeneous indel process, I divided the input MSA into ten blocks that differed in their indel model. The simulated blocks were then concatenated into a single alignment file and the GUIDANCE2 scores of the concatenated sequences were computed. Two different strategies were examined to account for a heterogeneous indel model across blocks. First, the values of 'IR' and 'a' were sampled from a normal distribution sampled around the values inferred by the OPTIMIM method (across the entire alignment) with SD of 0.2 around the mean (this simulation scenario is termed "Sampled"). Second, the 'IR' and 'a' values were estimated for each block separately, by the corresponding reference block (termed "Block-optimized"). As a control, the blocks were also simulated with the same values of 'IR' and 'a' parameters as inferred for the entire alignment, thus reflecting a homogenous model. In all three cases, the 'RL' parameter was set to the average length of the block's sequences.

As can be seen in Fig 19, the partitioning to blocks did not increase the complexity of the simulated MSAs compared to that of the homogenous model. In 11 out of 14 datasets, the difference was not significant (p-value > 0.05; ANOVA following Bonferroni correction). In the three datasets a significant deviation was observed (*18S* and *ITS* in the large dataset, and *26S* in the small dataset) the control simulations were not less complex than the other two simulation types (e.g., in the large *ITS* dataset, the simulation scenario with the highest GUIDANCE2 score was the one simulated under the ‘Sampled’ scenario).

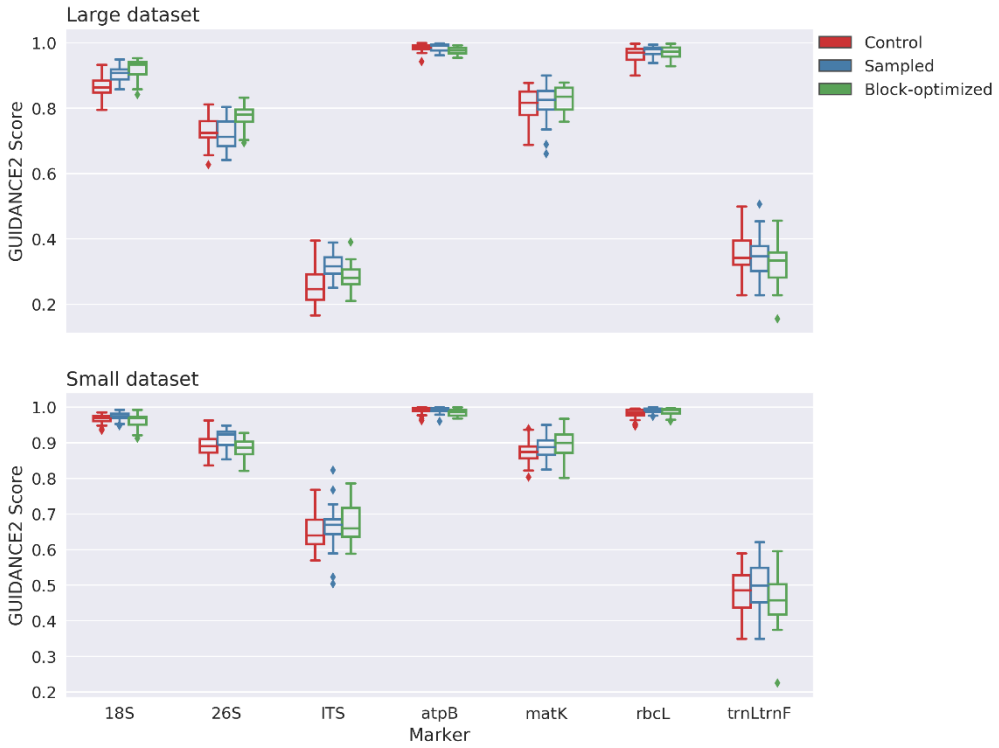


Fig 19. Alignment complexity of partitioned simulated data. Data was first partitioned into 10 blocks and then concatenated. For each marker, two boxplots of GUIDANCE2 scores of the concatenated sequences are presented, differ in the way the indel model was inferred for each partition, whether by sampling the parameters from a normal distribution (blue bars), or by inferring the parameters for each block separately (green). These boxplots are compared to the boxplot of the control (red), where the blocks were simulated with the same indel parameters inferred by OPTIMIM. Scores were obtained for 30 simulated MSAs. The size of the input sequences of the large dataset was limited to 1200.

3.3.3.3. Indel-length distribution

Another possible explanation for the simplified MSA produced via simulations may be a use of indel-length distribution that does not adequately describe the distribution observed in real biological data. As stated in the introduction, previous analyses of sequence data suggested that the length distribution of indels can be described by the power-law distribution, i.e., $p(x) = \frac{x^{-a}}{\varsigma(a)}$, where $p(x)$ is the frequency of indel of length x, a is the shape parameter, and $\varsigma(a)$ is the normalization factor.

Generally, the procedure of fitting a distribution to indel lengths is as follow. First, a representative set of sequence alignments is assembled. Second, the length of each indel should be extracted. Third, a distribution that fits the observed data should be chosen using

statistical methods of goodness of fit. Importantly, the second and the third steps of this procedure are not trivial, and the exact procedures undertaken in previous studies might have affect the derived conclusions. The first issue concerns with the definition of an indel. Most studies defined an indel as a gap block, and the indel-length distribution is constructed based on all gap blocks in the alignment. By doing so, however, all indels are considered as independent events (Fig 20). In reality, identical indels that are observed in multiple sequences may be the result of a single event, which occurred in an ancestral sequence during the process of evolution.

The second and perhaps more acute issue concerns with the method used to fit the power-law distribution to the data. Several previous studies deduced that the indel-length distribution follows the power-law distribution based on simple graphical examination (e.g., (Benner, Cohen and Gonnet, 1993; Zhang and Gerstein, 2003), which is certainly not an adequate statistical procedure. Others (e.g., Tao et al. 2007) used the Kolmogorov-Smirnov test (Massey Jr, 1951) and showed that the assumption that their sampled data were drawn from a power-law distribution cannot be rejected. However, as claimed by Clauset et al. (2009), obtaining a high *p*-value in such a statistical test does not necessarily mean that the power-law is the correct distribution for the data at hand. A possible explanation for the inability to reject the null hypothesis could be simply small sample size. Additionally, there may be other distributions that match the data equally or even better. Thus, we can rarely, if ever, be certain that an observed sample is drawn from a power-law distribution. The most we can say is that our observations are consistent with the hypothesis that the data are drawn from a power-law distribution. In some cases, we may also be able to rule out some other competing hypotheses.

Clauset et al. (2009) further presented a principled statistical framework for discerning whether a given sample follow the power-law distribution. Their suggested procedure follows these general steps. First, estimating the exponent parameter of the power-law distribution using maximum-likelihood optimization. Second, using a bootstrap strategy, the goodness-of-fit between the data and the inferred power-law distribution is computed. If the resulting *p*-value is greater than a defined threshold, the power law is considered a plausible hypothesis for the data, otherwise it is rejected (i.e., in this case, power-law is the null hypothesis and

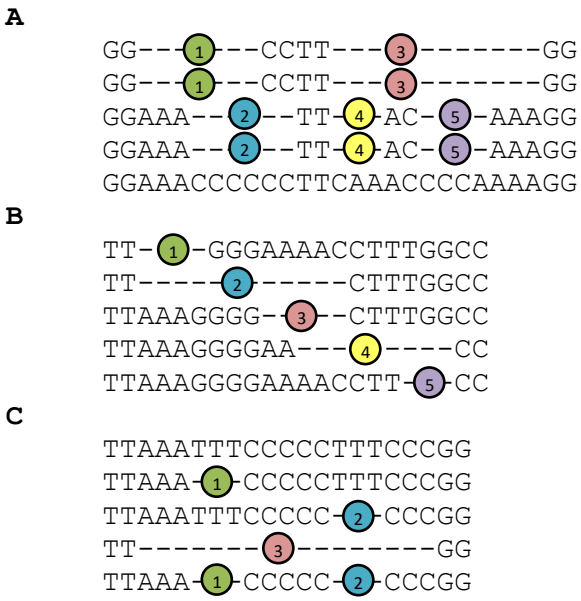


Fig 20. Fig SIC. Examples of simple gap coding in three alignments. In each alignment, the gap blocks are indicated by circled numbers. Gap blocks assigned as unique event are marked with the same number. Note the difference between the total number of gap blocks and the putative unique events. **(A)** There are 10 gap blocks and 5 indels events. **(B)** 5 gap blocks and 5 indels events. **(C)** 5 gap blocks and 3 indels events.

small *p*-values are indicative of poor fit). Third, to make a good case for or against the power-law for the data at hand, they suggested evaluating several other plausible distributions, such as exponential and log-normal. In particular, if the *p*-value obtained for the power-law distribution is higher than those of competing distributions, then there is a stronger case to believe that the sample is indeed power-law distributed. While the procedure detailed in Clauset et al. (2009) is widely referenced in the statistical literature, it was not applied thus far to evaluate indel-length distribution of biological data.

Thus, motivated by the question of whether the assumed indel distribution has an effect on the simulation complexity, I re-examined the assumption that indel lengths are drawn from the power-law distribution. To this end, I retrieved 1,000 protein MSAs of Euteleostomi (bony vertebrates) from the SELECTOME database (Proux et al. 2008), such that each MSA consists of at least 100 species, and at least 100 indel events (these cutoffs were made to ensure that the number of indels is sufficient). The indel events were determined using the "simple indel

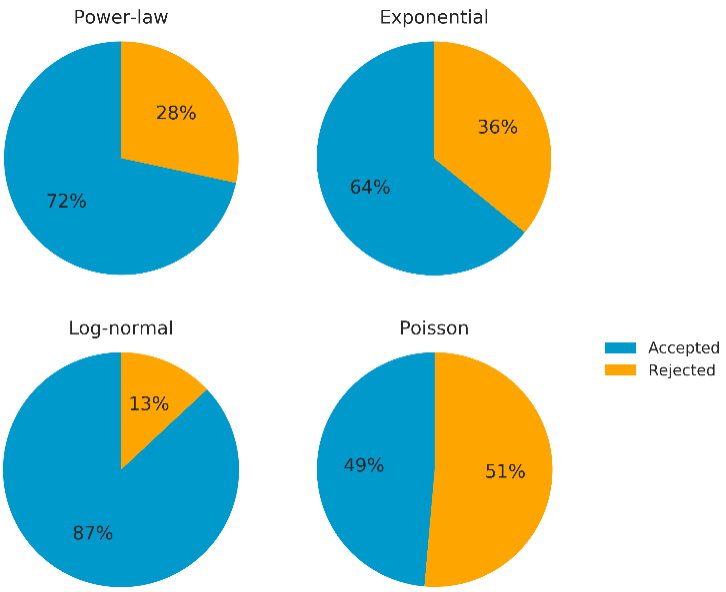


Fig 21. Pie chart of results of goodness-of-fit tests. For each distribution of the four following distributions: power-law, exponential, log-normal and Poisson, the corresponding pie shows the percentage of datasets that were accepted or rejected by the goodness-of-fit test.

coding" (SIC) method as described by Simmons & Ochoterena (2000). According to the SIC method, a unique indel is defined as gap block of the same size that are aligned to the same positions (Fig 20). Gaps at the edges of the sequences were ignored to avoid artificial indels as a result of partial sequence.

For each of the 1,000 MSAs examined, I calculated the goodness-of-fit between their indel-length distribution of unique indels and the power-law distribution, using the method of Clauset et al. (2009). As competing hypotheses, I used three other distributions: the exponential, log-normal, and Poisson distributions. My results revealed that for most datasets examined, multiple candidate distributions could not be rejected. Moreover, the power-law distribution was not the one for which most datasets could not be rejected, as the log-normal distribution seems to provide a plausible fit for a larger number of datasets (Fig 21). Additionally, for 18.3% of the datasets, all competing distributions provided a decent fit (i.e., they were not rejected by all four competing distributions), while for 0.7% of the datasets no distribution was found adequate (Fig 22).

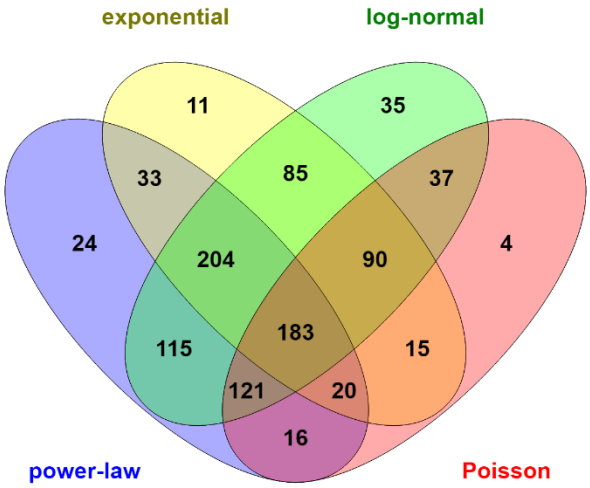


Fig 22. Venn diagram of datasets intersection. The diagrm shows the intersections of datasets for which the goddness-of-fit tests resulted in acception for at least one of the four following distributions: power-law, exponential, log-normal and Poisson. For example, the number in the intersection of all the ellipses indicates that 183 datasets were accepted by all distributions. The number beloning to the power-law alone indicaes that 24 datasets were accepted for power-law but not for the rest of the distributions. Total number of examined datasets: 993 out of 1000 (i.e, 7 datasets were not accepeted by all distributions).

As stated above, in order to claim that a given sample follows the power-law distribution, it is not enough to show that it is not rejected by the power-law, but rather that other distributions could be rejected. In my analysis, only 24 datasets, out of the 716 datasets that were accepted by the power-law, were also rejected by all other three competing distributions. This means that the 692 remaining datasets might as well be drawn from one or more of these distributions. In this case, Clauset et al (2009) suggested to compare the *p*-values obtained for each distribution. Doing so revealed that in these 692 datasets, 220 of the datasets with the highest *p*-value belonged to the power-law, meaning that in the rest of the cases the other accepted distributions fitted the data better (in 212 the exponential provided the best fit, in 389 the log-normal, and in 98 the Poisson).

Finally, I repeated the goodness-of-fit test for the united observed data – i.e., the lengths of each MSA were combined to one dataset in size of 192,681, with the aim to test which distribution provided the best fit to this one global distribution. there is one global

distribution. Using a similar testing procedure, the exponential distribution was rejected (*p*-value=0.0004), while the Poisson, power-law, and log-normal distributions were accepted, with a decreasing order of *p*-values, respectively (meaning that the data best fit the Poisson distribution) (*p*-values=0.81,0.74,0.26).

4. Discussion

The reconstruction of large phylogenies encompassing thousands of species allows for a broad perspective on the evolutionary relationships between species and is the key to the understanding of a variety of interesting questions. Currently, many mega phylogenies are available for use by the scientific community. Most of these phylogenies were inferred by the super-matrix approach, which uses a large MSA encompassing multiple loci and include all considered species. However, the accuracy level of phylogenies based on huge alignments and the amount of trust that should be given to them was not sufficiently investigated, especially in comparison to the ample research devoted to smaller phylogenies. The inclusion of many evolutionary distant species and the attempt to align their (sometimes too-far-apart) sequences may result in an erroneous alignment, and consequently in phylogenies of low accuracy. This could also result in errors in downstream analyses and to biased evolutionary theories based upon these phylogenies. Examining the quality of the inferred phylogenies is thus of great importance. This crucial point should also guide us on future improvements to phylogeny reconstruction methods.

The initial goal of this study was to evaluate the accuracy of existing large phylogenies inferred by current methods, with the hope to consequently develop an improved method for super-matrix reconstruction that will yield more accurate phylogenies. However, while working to achieve the first goal, which involved using sequence simulations, it was discovered that the data simulated with current approaches are by far too simplistic. This, in turn, did not allow me to evaluate the accuracy of mega-phylogenies inference. The research thus turned to a new direction, in which I studied the complexity of current simulation approaches. Currently, there exists several sequence simulators that are commonly used by the community. However, these simulators produce unrealistic sequences, and the evolutionary model parameters they require to specify are difficult to estimate. Simulations play a major role in many biological fields and are sometimes the only way to evaluate computational performances, thus improving the realism of sequences generated by current simulators is of great necessity.

In the first part of this study, I evaluated the accuracy of large inferred phylogenies either by comparing to taxonomical knowledge or by using simulations to test how their accuracy is affected by alignment complexity. I found that the degree of agreement between the examined phylogenies and the taxonomical knowledge is not as high as might be expected, and that the more complex the alignment, the less accurate the inferred phylogenies (see Results, section 3.2). However, the effect of the alignment complexity on phylogenetic accuracy was expected to be more substantial, i.e., complex alignments reduced the accuracy less than expected. One possible explanation to this finding is that the tree reconstruction methodology is the cause for most of the inaccuracies in large phylogenies. A more likely possibility, which I chose to focus on, was that the simulated data do not resemble the biological sequences in terms of alignment complexity, and thus the number and variety of the sequences did not affect much the inference of the alignments as in biological data. Indeed, I found that the complexity of the simulated sequences is substantially lower than that found in the set of empirical data studied here (see Results, section 3.3.2).

The second part of this study was guided by the question of whether we simulate realistic sequences that resemble real biological data. Two central issues were investigated: determining the evolutionary model parameters, and evaluating the complexity of both the reference and the simulated alignments.

First, I discuss the issue of estimating the parameters, particularly the indel parameters, under which the simulation is performed. Notably, despite the great impact of these parameters on the resulting simulations, I am aware of only three studies devoted to the task of determining their values (Cartwright, 2005; Levy Karin *et al.*, 2015; Levy Karin *et al.*, 2017). Furthermore, I found that the parameters inferred by existing tools fail to generate alignments that resemble the reference alignments in two very basic characteristics: the length of the alignment and the average number of gap blocks, thus rendering them practically useless. As a result, I presented a new procedure, termed OPTIMIM, to infer parameter values that resulted in alignments that successfully resembled these characteristics. However, although these statistics are of utmost important, basing the inference on these alone is insufficient to determine similarity to the input biological data. Many other characteristics such as the scattering of the gap blocks along the alignment, the distribution of their lengths, and the

issue of gaps at the edges of the alignment, should be considered. Many such statistics are indeed considered by SpartaABC (Levy Karin *et al.*, 2017), but for unknown reason its results were poor. A possible explanation is that the performance of this tool was never reported on real biological data, but only on simulated ones. Moreover, the choice of the parameters did not substantially affect the alignment complexity. Finally, it seems that this research field has not received enough attention. The little amount of citations to the existing tools shows that apparently many of the studies that simulated sequences used either arbitrary parameters or local solutions for the indel model.

In this study, alignment complexity was quantitively measured by the "GUIDANCE2 score", which is the average reliability scores of the alignment columns provided by GUIDANCE2 (Sela *et al.*, 2015). Originally, the GUIDANCE2 tool was developed in the context of identifying unreliable columns (or positions) in sequence alignments, so that one could consider filtering these columns before subsequent analyses. This score does not measure the quality of the alignment, but rather its robustness. Here I interpreted robustness as complexity – the less the alignment is robust to perturbations in the alignment parameters and guide tree, the more complex its computation. GUIDANCE2 was one of several considered tools for measuring alignment reliability. The clear and easy-to-interpret results obtained by this tool, together with its popularity (for measuring alignment reliability) and ease of use, have made GUIDANCE2 a preferred and sufficient candidate to demonstrate the claim regarding the simplistic simulations. Notably, GUIDANCE2 can be used for the purpose of measuring alignment complexity in a more delicate way, for example, by examining the distribution of the column's reliability scores instead of their total average, and thus achieving a more detailed picture of the alignment complexity.

The low GUIDANCE2 scores obtained for most of the biological sequences were interpreted as indication for a complex and diverged set of sequences, in contrast to the easy-to-align sequences produced using simulations. Notably, these scores differed between the seven markers analyzed: the scores of *26S*, *ITS*, and *trnL-F* (and *18S* in the large dataset) were much lower than those of *atpB*, *matK*, and *rbcL*. This difference may be explained by the difference in the rate of evolution between the markers. Rapidly evolving sequences are more challenging to align than slowly evolving sequences, which leads to less confidence in the

alignment and thus lower GUIDANCE2 scores. Indeed, *atpB*, *matK*, and *rbcL* are protein-coding genes and their higher GUIDANCE2 scores can be explained by their low rate of evolution compared to the other markers analyzed here. Another very likely possibility, which is discussed below, concerns the variation of sequence lengths, which is specifically high for the *26S* and *ITS* markers. Two exceptions were observed for the *atpB* and *rbcL* markers, whose GUIDANCE2 scores obtained for their reference alignments (of the small datasets) were very high. This is probably the result of the very small numbers of gap blocks observed in their corresponding MSAs. This can also explain why in these cases the GUIDANCE2 scores obtained for the alignments simulated under SpartaABC-inferred parameters were notably lower: their average numbers of gap blocks were much higher than these of the reference ones. In contrast, when the OPTIMIM method was used to infer indel parameters, the resulting simulated alignments obtained scores similar to those obtained for the reference MSAs.

Evidently, GUIDANCE2 scores of the large dataset (obtained for both the reference and the simulated sequences) were overall lower than those of the small dataset. This is expected since the larger the dataset, the larger the diversity of the sequences, and thus the lower the GUIDANCE2 scores. The effect of adding more sequences was particularly noticeable for *ITS* and *26S,* whose GUIDANCE2 scores of the large dataset (0.003 and 0.00003, respectively), were substantially lower compared to those obtained for the small dataset (0.05 and 0.13, respectively).

I examined several factors which may affect the alignment complexity: the distribution of the sequence lengths, the use of non-homogeneous indel model, and the distribution of the indel lengths. Starting from the first and probably the most important one, I examined the sequence lengths distribution and found that the variance of the simulated sequence length is very small, as was clearly illustrated by the dominant mode shown in their histograms (Fig 17). In contrast, the variance observed in the reference sequences was much higher. The variability in the sequence lengths could potentially stems from two sources: gap blocks at the edges of the alignment and internal gap blocks. By cutting the edges of the simulated sequences at the same proportion as in the reference sequences, the variance of the simulated sequences became similar to that of the reference, meaning that the main source of the gap blocks was due to gaps at the edges. Consequently, the procedure of cutting the edges resulted in more

complex alignments as measured by lower GUIDANCE2 scores. The increase in complexity was respective to the magnitude of the variance, such that scores of markers with high length variance, such as *26S* and *ITS*, decreased more than markers with low variance. It can thus be concluded that the inclusion of short sequences in the alignment reduces its accuracy. This may be expected since shorter sequences obviously include more gap blocks, and thus harder to align. However, it could also be expected that such gaps will appear in a single block at the beginning (or end) of the sequence. Evidently, this was not the case as the placement of gaps was dependent on the exact parameters and guide tree used in the alignment procedure (and thus led to low GUIDANCE2 scores). As mentioned in the Results, short sequences are mostly caused due to partial sequencing, rather by biological forces. This technical obstacle is likely to be solved over time. However, the fact that the markers that were not sequenced to their whole length are also fast-evolving (e.g., *26S*, *ITS*), did not allow me to fully separate the effect of these two factors on alignment complexity. Thus, it will be important to evaluate the complexity of fast-evolving markers which are entirely sequenced (for example, by repeating the analysis while filtering partial sequences).

The second factor I examined as a possible effect on alignment complexity was the use of non-homogenous indel model. Non-homogenous indel models, in contrast to simple indel models, do not assume that the evolutionary pattern is the same along all branches or sequence sites (or both). Inferring an indel model for each branch separately cannot be achieved by current tools, so I focused on inferring different indel model for different sites (columns in the alignment). This was done by an arbitrary partitioning of the reference alignment to ten blocks of equal size differ in the indel model, simulating each block with its own model, and combining them to one alignment. However, the complexity of the combined alignment was not increased substantially. Two possibilities may explain this result: either partitioning does not have the desirable effect of producing more complicated alignments, or the division should be done with some biological logic, rather than performed in a fixed-sized manner.

The third and last factor I examined as a possible effect on alignment complexity is the distribution of the indel lengths. I started by reexamining the prevailing assumption that indel lengths follow the power-law distribution (Benner, Cohen and Gonnet, 1993; Gu and Li, 1995; Zhang and Gerstein, 2003; Chang and Benner, 2004; Yamane, Yano and Kawahara, 2006;

Cartwright, 2008). However, Clauset et al (2009) have claimed that determining that a particular sample follows the power-law distribution could rarely be stated with certainty. This uncertainty is caused mainly due to lack of data in the tail of the distribution, which represents large but rare events. This issue is especially true in the case of indel lengths, where long indels are indeed rare. Notably, long indels can be found at the edges of the alignment, but they are probably the result of partial sequencing and thus should not be considered. using the goodness-of-fit test suggested by Clauset et al (2009), I found that although the power-law could not be rejected in 72% of the examined datasets, only 4% of these cases were also rejected by the three other examined distributions. Additionally, for the log-normal distribution, the percentage of non-rejected datasets was even higher. This indicates that in many cases the power-law does not provide the only possible alternative. To increase the power of the test, I repeated it for a data that combined all individual datasets together. On this sample, I found that the Poisson distribution was favored, although the power-law did not lag far behind (only the exponential distribution was rejected, meaning all the three others are possible). Taking together, the results suggest that the power-law may not be the most suitable one to describe the distribution of indel lengths. Moreover, it is likely that there is no single distribution that provides the best fit for all MSA and perhaps a more flexible distribution should be evaluated. Notwithstanding, it is yet to be examined whether alternative indel-length distributions would affect the alignment complexity. To this end, simulated data should be generated with alternative distributions. Unfortunately, there is no current method to estimate the parameters of such distributions (e.g., as done by SpartaABC for power-law).

Finally, a number of potential caveats of this study should be noted. First, my main set of analyses (aside those concerning the fit of the power law to the indel-length distribution) were based on only seven genomic markers. Although the difference between the complexity of biological and simulated alignments of these markers was clearly shown, additional datasets are needed to strengthen this finding. Another matter concerns the measure used to evaluate the alignment complexity. My analyses relied on GUIDANCE2 scores, which practically evaluates how perturbations in the parameters that govern the alignment computation affects the resulting MSA. It would be interesting to examine whether alternative alignment reliability measures could reveal other aspects of this process. Lastly, although being a popular simulator, results are solely based on the INDELible simulator. Recently, a more flexible

simulator termed PhyloSim (Sipos *et al.*, 2011) was presented and it would be interesting to evaluate the complexity of the simulated sequences it produces.

In conclusion, the main contribution of this study is the understating that the field of sequence simulations is still in its infancy, resulting in sequences that are not realistic enough. Sequence simulations are usually used as a means to valid the development of new tools and to compare performances among several alternatives methodologies for the same task. Thus, the development of more sophisticated simulation tools is usually not regarded as a central scientific goal. Yet, the use of more realistic simulations should have a dramatic effect on the many analyses in the filed on molecular evolution and it is vital that such simulations would reflect the biological data at hand, or otherwise, the conclusion of much research could be greatly biased.

5. Methods

5.1. Index of main programs used in this study

ETE Toolkit (http://etetoolkit.org/) -A Python framework to work with trees which also provides utilities to query the NCBI Taxonomy database.

ExaML (Kozlov, Aberer and Stamatakis, 2015) - A very fast phylogeny reconstruction method that is suitable for extremely large-scale phylogenetic inference.

INDELible (Fletcher and Yang, 2009) - A sequence simulator program.

Ktreedist (Soria-Carrasco *et al.*, 2007)- A software for calculating distance between phylogenies.

Lambda (Cartwright, 2005) – A script for indel model inference implemented as part of the Dawg package.

MAFFT (Katoh and Standley, 2013) **-** A multiple-alignment program for amino acid or nucleotide sequences.

PHLAWD (Smith, Beaulieu and Donoghue, 2009) - A software for the assembly of super-matrices that can be used for mega-phylogeny inference. The software retrieves GenBank sequences of a clade of interest and subsequently constructs its multiple alignment.

RAxML (Stamatakis and Aberer, 2013) - A software for maximum likelihood based inference of phylogenetic trees.

SPARTA (Levy Karin *et al.*, 2015) - A software for indel model inference.

SpartaABC (Levy Karin *et al.*, 2017) - A software for indel model inference.

5.2. Data assembly

I compiled nucleotide sequence data from GenBank for the Magnoliophyta (flowering plants). Seven markers were selected, which are the ones used to reconstruct the phylogeny of Zanne et al. (2014) choice: *18S* rDNA, *26S* rDNA, *ITS*, *matK*, *rbcL*, *atpB*, and *trnL-F*. These seven nucleotide markers are among the most commonly sampled regions used in molecular plant systematic studies and were originally chosen to include both slowly evolving regions that have been broadly sampled across the clade (e.g., *rbcL*, *18S* rDNA) and more rapidly evolving regions that have been densely sampled for species-level phylogenetic studies (e.g., *ITS*, *trnL-F*).

Data from GenBank for these seven markers were retrieved and cleaned using the PHLAWD pipeline (v.3.3a), resulting in a data set that included 175,060 sequences representing over 100,000 species. To maximize coverage and reduce computational complexity, I reduced the species data according to markers coverage to form the two following data sets, which will be referred to as the reference datasets (Table 5):

a) Partial coverage: species for which at least 4 markers were available. This will be referred to as **large dataset** - resulted in 5,113 species representing over 60 orders.
b) Full coverage: species for which all 7 markers were available. This will be referred to as **small dataset** - resulted in 132 species representing over 40 orders.

Multiple sequence alignments were generated with MAFFT (v.6.864b) using the auto option (chosen strategy: FFT-NS-2). Individual alignments were then concatenated to form a single data matrix (i.e., "super-matrix"). Table 5 presents the number of species and the length of the alignments that were generated for each marker, as well as the number of species and length of the concatenated MSAs.

Overall, the concatenated matrix of the large data set had a large proportion (0.64) of missing data, although proportions varied among the seven gene regions. I report them as per-site and per-taxon proportion of missing data, respectively: *18S* (0.10, 0.66), *26S* (0.46, 0.75), *ITS* (0.72, 0.12), *atpB* (0.07, 0.49), *matK* (0.29, 0.30), *rbcL* (0.09, 0.04) and *trnL-F* (0.58, 0.20).

Table 5. Number of species and length of individual and concatenated MSAs of reference sequences data sets.

Marker	Large data set		Small data set	
	Number of species	MSA length	Number of species	MSA length
18S	1,725	1,836	132	1,813
26S	1,289	3,715	132	3,494
ITS	4,483	2,205	132	1,041
atpB	2,604	1,521	132	1,517
matK	3,585	2,164	132	1,891
rbcL	4,931	1,433	132	1,427
trnL-F	4,107	1,938	132	1,326
Concatenated	5,113	14,812	132	12,509

Finally, I performed maximum-likelihood-based phylogenetic analyses of the resulting super-matrix using EXaML (v.3.0.17), partitioned by gene region. The phylogeny based on the large data is the reference phylogeny referred to in the Results (see Introduction, section 3.2). The maximum-likelihood-based phylogenetic analyses were also performed for each of the individual alignments, resulting in seven phylogenies that are used in the simulations (see section 5.5).

5.3. Assessing monophyly in mega phylogeny

This step involved the assessment of the monophyly level of a group of interest within a given phylogeny. Three monophyly scores were computed, referred to as strict, score 1 and score 2 (Fig 12). Those scores were computed for a published phylogeny reconstructed by Zanne et al. (2014) and two large phylogenies reconstructed during this study (termed PHLAWD and MAFFT phylogenies). The following are details of these procedures:

Calculating monophyly scores - To calculate the monophyly scores I wrote a Python script that uses the ETE module. The ETE module offers functions, such getting the most common ancestor of given taxa, and extracting the species of a given clade and rank in the NCBI Taxonomy database.

Reconstruction of Zanne's phylogeny - The phylogeny of Zanne et al. (2014) was reconstructed based on seven markers above mentioned. Genetic data were compiled and aligned using the PHLAWD pipeline and maximum-likelihood-based phylogenetic analyses of the resulting alignment were performed using RAxML (v.7.4.1), partitioned by gene region and with major clades (that is, families and orders) constrained according to the APG III (Group, 2009) classification system.

Reconstruction of the PHLAWD and MAFFT phylogenies - The phylogenies were reconstructed using the same seven nucleotide markers as in Zanne et al. Sequence data for each marker were retrieved as described in section 5.2. The exemplar of each marker, which should be supplied to PHLAWD, was as used Zanne et al. (2014). Sequences of species that were not found in Zann's tree were removed from the MSA produced by PHLAWD. Because PHLAWD incorporate taxonomic constraints as part of the alignment procedure, the retrieved sequences were additionally re-aligned with MAFFT (v.6.864b) (with the default options). Maximum-likelihood-based phylogenetic analyses of the two alignments were performed using ExaML (v.3.0.17) using the GTR+ Γ model of nucleotide substitution, partitioned by marker.

5.4. Testing the effect of alignment complexity on the accuracy of inferred phylogenies

This section specifies in more details the simulation approach that was used to test the hypothesis that the inclusion of highly divergent sequences within the super-matrix approach would lead to poor alignments and inaccurate trees. First, I assembled an MSA for a diverse set of ~5,000 flowering plants species for seven genomic detailed above (see section 5.2). This MSA was used to reconstruct a phylogeny, which was referred to as the reference phylogeny (see Results, section 3.2). Second, based on the reference phylogeny, I simulated sequences with the aim to generate sequences that mimic the real data, using the INDELible simulator (see section 5.5). Third, I iteratively extracted from the reference phylogeny monophyletic groups of different sizes, $g_1,g_2,...,g_n$, such that smaller groups are contained within the larger ones, resulting in increasingly larger phylogenies $t_1,t_2,...,t_n$. For each species group g_i, I extracted the corresponding "true" sequence alignment from the simulated data, and used these to reconstruct a set of inferred trees $s_1,s_2,...,s_n$. The final step was to compute the distances between t_i and $s_i,s_{i+1},... ,s_n$ using the Ktreedist distance (Soria-Carrasco *et al.*, 2007).

When comparing trees of different sizes (i.e., t_i to s_{i+1}) I pruned from the larger tree species that were not included in the smaller one.

5.5. Simulating data

Simulations were conducted using the INDELible simulator (V1.03). The INDELible simulator requires the specification of substitution and indel models. As a substitution model, I used the GTR+Γ model, whose rate parameters were set to those inferred for each marker using ExaML during the phylogeny reconstruction step. To estimate the three indel parameters (termed 'RL', 'IR', and 'a'; defined in Results, section 3.3.1) from the reference alignments (for each marker separately), three methods were used:

a) Lambda - the inputs were MSA and a phylogeny of the analyzed marker alignment.
b) SPARTA - required the same inputs as lambda.
c) SpartaABC - in addition to the inputs required by Lambda and SPARTA, SpartaABC also requires the minimal and maximal value of 'IR' and 'RL' parameters, which define the search range (see Introduction, section 1.9). These values were set to 0 and 0.01, respectively, based on the assumption that an indel rate higher than 0.01 is unrealistic. The min and max 'RL' values were set to the minimal and maximal sequence length in the input alignment. The number of iterations was set to 100,000. The output of SpartaABC is a list of indel parameters values sorted by the probability to yield the input MSA. The inferred indel parameters were chosen as the average of the 50 top sets, as recommended by the authors. Both Sparta and SpartaABC run INDELible as their simulator.

Indel-length distribution was set as the power-law distribution. Since the upper range of the power-law is infinite, INDELible offers an option to set a maximal indel length. In most simulations in this study this value was set to 50. For the simulations described in section 5.4 the branches of the input phylogenies were multiplied by two in order to increase the divergence among the output sequences, and the maximal indel length was set to 20.

6. References

Benner, S. A., Cohen, M. A. and Gonnet, G. H. (1993) 'Empirical and structural models for insertions and deletions in the divergent evolution of proteins', *Journal of molecular biology*. Elsevier, 229(4), pp. 1065–1082.

Blackburne, B. P. and Whelan, S. (2012) 'Class of multiple sequence alignment algorithm affects genomic analysis', *Molecular biology and evolution*. Oxford University Press, 30(3), pp. 642–653.

Blanchette, M. *et al.* (2004) 'Reconstructing large regions of an ancestral mammalian genome in silico', *Genome research*. Cold Spring Harbor Lab, 14(12), pp. 2412–2423.

Bull, J. J. *et al.* (1993) 'Experimental molecular evolution of bacteriophage T7', *Evolution*. Wiley Online Library, 47(4), pp. 993–1007.

Cartwright, R. A. (2005) 'DNA assembly with gaps (Dawg): simulating sequence evolution', *Bioinformatics*. Oxford University Press, 21(Suppl_3), pp. iii31-iii38.

Cartwright, R. A. (2008) 'Problems and solutions for estimating indel rates and length distributions', *Molecular biology and evolution*. Oxford University Press, 26(2), pp. 473–480.

Chang, M. S. S. and Benner, S. A. (2004) 'Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments', *Journal of molecular biology*. Elsevier, 341(2), pp. 617–631.

Chor, B. and Tuller, T. (2005) 'Maximum likelihood of evolutionary trees: hardness and approximation', *Bioinformatics*. Oxford University Press, 21(suppl_1), pp. i97--i106.

Christenhusz, M. J. M. and Byng, J. W. (2016) 'The number of known plants species in the world and its annual increase. Phytotaxa 261 (3): 201–217'.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009) 'Power-law distributions in empirical data', *SIAM review*. SIAM, 51(4), pp. 661–703.

Crowder, L. B. *et al.* (2006) 'Resolving mismatches in US ocean governance', *SCIENCE-NEW YORK THEN WASHINGTON-*. American Association for the Advancement of Science, 313(5787), p. 617.

Durbin, R. *et al.* (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Edgar, R. C. and Batzoglou, S. (2006) 'Multiple sequence alignment', *Current opinion in structural biology*. Elsevier, 16(3), pp. 368–373.

Felsenstein, J. (1981) 'Evolutionary trees from DNA sequences: a maximum likelihood approach', *Journal of molecular evolution*. Springer, 17(6), pp. 368–376.

Felsenstein, J. (1985) 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution*. Wiley Online Library, 39(4), pp. 783–791.

Fletcher, W. and Yang, Z. (2009) 'INDELible: a flexible simulator of biological sequence evolution', *Molecular biology and evolution*. Oxford University Press, 26(8), pp. 1879–1888.

Gaut, B. S. and Lewis, P. O. (1995) 'Success of maximum likelihood phylogeny inference in the four-taxon case.', *Molecular Biology and Evolution*, 12(1), pp. 152–162.

Gillespie, D. T. (1977) 'Exact stochastic simulation of coupled chemical reactions', *The journal of physical chemistry*. ACS Publications, 81(25), pp. 2340–2361.

Goldman, N. (1993) 'Simple diagnostic statistical tests of models for DNA substitution', *Journal of Molecular Evolution*. Springer, 37(6), pp. 650–661.

Group, A. P. (2009) 'An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III', *Botanical journal of the Linnean Society*. Oxford University Press, 161(2), pp. 105–121.

Gu, X. and Li, W.-H. (1995) 'The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment', *Journal of molecular evolution*. Springer, 40(4), pp. 464–473.

Guindon, S. and Gascuel, O. (2003) 'A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood', *Systematic biology*. Society of Systematic Zoology, 52(5), pp. 696–704.

Hickson, R. E., Simon, C. and Perrey, S. W. (2000) 'The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence', *Molecular Biology and Evolution*. Oxford University Press, 17(4), pp. 530–539.

Hillis, D. M. *et al.* (1992) 'Experimental phylogenetics: generation of a known phylogeny', *Science*. American Association for the Advancement of Science, 255(5044), pp. 589–592.

Hinchliff, C. E. and Smith, S. A. (2014) 'Some limitations of public sequence data for phylogenetic inference (in plants)', *PLoS One*. Public Library of Science, 9(7), p. e98986.

Huang, W., Umbach, D. M. and Li, L. (2005) 'Accurate anchoring alignment of divergent sequences', *Bioinformatics*. Oxford University Press, 22(1), pp. 29–34.

Huelsenbeck, J. P. (1995) 'Performance of phylogenetic methods in simulation', *Systematic biology*. Society of Systematic Biologists, 44(1), pp. 17–48.

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics*. Oxford University Press, 17(8), pp. 754–755.

Jetz, W. *et al.* (2012) 'The global diversity of birds in space and time', *Nature*. Nature Publishing Group, 491(7424), p. 444.

De Jong, H. (2002) 'Modeling and simulation of genetic regulatory systems: a literature review', *Journal of computational biology*. Mary Ann Liebert, Inc., 9(1), pp. 67–103.

Jukes, T. H. and Cantor, C. R. (1969) 'Evolution of protein molecules', *Mammalian protein metabolism*. New York, 3(21), p. 132.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Katoh, K. and Standley, D. M. (2016) 'A simple method to control over-alignment in the MAFFT

multiple sequence alignment program', *Bioinformatics*. Oxford University Press, 32(13), pp. 1933–1942.

Kemena, C. and Notredame, C. (2009) 'Upcoming challenges for multiple sequence alignment methods in the high-throughput era', *Bioinformatics*. Oxford University Press, 25(19), pp. 2455–2465.

Kozlov, A. M., Aberer, A. J. and Stamatakis, A. (2015) 'ExaML version 3: A tool for phylogenomic analyses on supercomputers', *Bioinformatics*, 31(15), pp. 2577–2579. doi: 10.1093/bioinformatics/btv184.

Kuhner, M. K. and Felsenstein, J. (1994) 'A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.', *Molecular biology and evolution*, 11(3), pp. 459–468.

Kumar, S. and Filipski, A. (2007) 'Multiple sequence alignment: in pursuit of homologous DNA positions', *Genome research*. Cold Spring Harbor Lab, 17(2), pp. 127–135.

Law, A. M., Kelton, W. D. and Kelton, W. D. (1991) *Simulation modeling and analysis*. McGraw-Hill New York.

Levy Karin, E. *et al.* (2015) 'Inferring indel parameters using a simulation-based approach', *Genome biology and evolution*. Oxford University Press, 7(12), pp. 3226–3238.

Levy Karin, E. *et al.* (2017) 'Inferring rates and length-distributions of indels using approximate Bayesian computation', *Genome biology and evolution*. Oxford University Press, 9(5), pp. 1280–1294.

Liu, J. *et al.* (2011) 'Stable nanosecond pulse generation from a graphene-based passively Q-switched Yb-doped fiber laser', *Optics letters*. Optical Society of America, 36(20), pp. 4008–4010.

Loew, L. M. and Schaff, J. C. (2001) 'The Virtual Cell: a software environment for computational cell biology', *TRENDS in Biotechnology*. Elsevier, 19(10), pp. 401–406.

Massey Jr, F. J. (1951) 'The Kolmogorov-Smirnov test for goodness of fit', *Journal of the American statistical Association*. Taylor & Francis Group, 46(253), pp. 68–78.

Needleman, S. B. and Wunsch, C. D. (1970) 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of molecular biology*. Elsevier, 48(3), pp. 443–453.

Notredame, C. (2007) 'Recent evolutions of multiple sequence alignment algorithms', *PLoS computational biology*. Public Library of Science, 3(8), p. e123.

Nuin, P. A. S., Wang, Z. and Tillier, E. R. M. (2006) 'The accuracy of several multiple sequence alignment programs for proteins', *BMC bioinformatics*. BioMed Central, 7(1), p. 471.

Ogden, T. H. and Rosenberg, M. S. (2006) 'Multiple sequence alignment accuracy and phylogenetic inference', *Systematic biology*. Society of Systematic Zoology, 55(2), pp. 314–328.

Pyron, R. A., Burbrink, F. T. and Wiens, J. J. (2013) 'A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes', *BMC evolutionary biology*. BioMed Central, 13(1), p. 93.

de Queiroz, A. and Gatesy, J. (2007) 'The supermatrix approach to systematics', *Trends in ecology & evolution*. Elsevier, 22(1), pp. 34–41.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing

phylogenetic trees.', *Molecular biology and evolution*, 4(4), pp. 406–425.

Sayers, E. W. (2009) 'Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E', *Database resources of the National Center for Biotechnology Information. Nucl Acids Res*, 37, pp. D5--D15.

Sela, I. *et al.* (2015) 'GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters', *Nucleic acids research*. Oxford University Press, 43(W1), pp. W7–W14.

Simmons, M. P. and Ochoterena, H. (2000) 'Gaps as characters in sequence-based phylogenetic analyses', *Systematic biology*. JSTOR, 49(2), pp. 369–381.

Sipos, B. *et al.* (2011) 'PhyloSim-Monte Carlo simulation of sequence evolution in the R statistical computing environment', *BMC bioinformatics*. BioMed Central, 12(1), p. 104.

Slepchenko, B. M. *et al.* (2002) 'Computational cell biology: spatiotemporal simulation of cellular events', *Annual review of biophysics and biomolecular structure*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 31(1), pp. 423–441.

Smith, S. A., Beaulieu, J. M. and Donoghue, M. J. (2009) 'Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches', *BMC evolutionary biology*. BioMed Central, 9(1), p. 37.

Smith, S. A. and Brown, J. W. (2018) 'Constructing a broadly inclusive seed plant phylogeny', *American journal of botany*. Wiley Online Library, 105(3), pp. 302–314.

Soria-Carrasco, V. *et al.* (2007) 'The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees', *Bioinformatics*. Oxford University Press, 23(21), pp. 2954–2956.

Stamatakis, A. and Aberer, A. J. (2013) 'Novel parallelization schemes for large-scale likelihood-based phylogenetic inference', in *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium On*. IEEE, pp. 1195–1204.

Strope, C. L. *et al.* (2009) 'Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0', *Molecular biology and evolution*. Oxford University Press, 26(11), pp. 2581–2593.

Swenson, N. G. (2009) 'Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities', *PloS one*. Public Library of Science, 4(2), p. e4390.

Tao, S. *et al.* (2007) 'Patterns of insertion and deletion in mammalian genomes', *Current genomics*. Bentham Science Publishers, 8(6), pp. 370–378.

Tavaré, S. (1986) 'Some probabilistic and statistical problems in the analysis of DNA sequences', *Lectures on mathematics in the life sciences*, 17(2), pp. 57–86.

Thompson, J. D. *et al.* (2005) 'BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark', *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 61(1), pp. 127–136.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) 'CLUSTAL W: improving the sensitivity of

progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic acids research*. Oxford university press, 22(22), pp. 4673–4680.

Vinga, S. and Almeida, J. (2003) 'Alignment-free sequence comparison—a review', *Bioinformatics*. Oxford University Press, 19(4), pp. 513–523.

Van Walle, I., Lasters, I. and Wyns, L. (2004) 'Align-m—a new algorithm for multiple alignment of highly divergent sequences', *Bioinformatics*. Oxford University Press, 20(9), pp. 1428–1435.

Wang, L. and Jiang, T. (1994) 'On the complexity of multiple sequence alignment', *Journal of computational biology*, 1(4), pp. 337–348.

Whelan, S. and Goldman, N. (2001) 'A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach', *Molecular biology and evolution*. Oxford University Press, 18(5), pp. 691–699.

Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. (2008) 'Alignment uncertainty and genomic analysis', *Science*. American Association for the Advancement of Science, 319(5862), pp. 473–476.

Worobey, M., Han, G.-Z. and Rambaut, A. (2014) 'A synchronized global sweep of the internal genes of modern avian influenza virus', *Nature*. NIH Public Access, 508(7495), p. 254.

Xia, X. (2016) 'PhyPA: phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences', *Molecular phylogenetics and evolution*. Elsevier, 102, pp. 331–343.

Yamane, K., Yano, K. and Kawahara, T. (2006) 'Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice', *DNA research*. Oxford University Press, 13(5), pp. 197–204.

Zanne, A. E. *et al.* (2014) 'Three keys to the radiation of angiosperms into freezing environments'. Nature Publishing Group.

Zapata, F. *et al.* (2015) 'Phylogenomic analyses support traditional relationships within Cnidaria', *PLoS One*. Public Library of Science, 10(10), p. e0139068.

Zhang, Z. and Gerstein, M. (2003) 'Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes', *Nucleic acids research*. Oxford University Press, 31(18), pp. 5338–5348.

אוניברסיטת תל-אביב
הפקולטה למדעי החיים ע"ש ג'ורג' ס. וויז
המדרשה לתארים מתקדמים

לקראת סימולציות רצפים ריאליסטיות
למטרת בניית עצים פילוגנטיים גדולים

עבודה זו הוגשה לקראת התואר "מוסמך אוניברסיטה"

במסלול לביו-אינפורמטיקה באוניברסיטת תל-אביב

על-ידי

נעמי הדר

העבודה הוכנה במחלקה לביולוגיה מולקולרית ואקולוגיה של צמחים

של אוניברסיטת תל-אביב

בהנחיית

פרופ' איתי מירוז

תמוז התשע"ח

61

תקציר

השחזור של עצים פילוגנטיים גדולים הכוללים אלפי מינים מאפשר לנו מבט רחב על תהליכים אבולוציוניים ומשמש נקודת מוצא לחקר תופעות אבולוציוניות רבות. שיטות לשחזור פילוגנטי מבוססות על עימוד (alignment) מרובה רצפים (MSA), הכולל מספר לוקוסים. המידה שבה עץ פילוגנטי משקף את ההיסטוריה האמתית תלויה באופן ישיר בדיוק של הMSA. אולם, המשימה של עימוד רצפים הופכת למאתגרת יותר ככל שהרצפים באנליזה בעלי שונות רבה יותר. לפיכך, הבנייה של MSA גדול המכיל רצפים שנדגמו ממגוון רחב של מינים עלול להוביל לתוצאות עימוד עלובות ולעצים פילוגנטיים לא מדויקים. במחקר זה, מטרתי הראשונית היא להעריך את השפעת המורכבות של הMSA על דיוק העצים הפילוגנטיים המוסקים באמצעות סימולציית רצפים. בפרט, חקרתי את ההשערה שככל שהעימוד מורכב יותר כך הפילוגנזה שתתקבל תהייה מדויקת פחות. השפעה זו אכן נצפתה; עם זאת, היא הייתה נמוכה מהצפוי. החשד היה שסימולציית הרצפים אינן מציאותיות מספיק, מה שהופך אותן ללא אינפורמטיביות על מנת ללמוד את הדיוק של עצים פילוגנטיים ענקיים. המחקר שלי לפיכך התמקד בזיהוי החסרונות של גישות הסימולציה הנוכחיות ובמסלולים שבאמצעותם ניתן לשפר את סימולציית הרצפים. ראשית, מצאתי כי השיטות הנוכחיות להסקת הפרמטרים ששולטים בדינמיקה של אירועי מחיקה והכנסה (indels) נוחלים כישלון בייצור MSAs שמזכירים את ה- MSA המקורי. כדי להסיק את הפרמטרים האלה נדרשתי לפתח פרוצדורה חדשה. שנית, באמצעות שימוש בכלים של אמינות של MSAs, גיליתי שה-MSAs המסומלצים שונים באופן דרסטי במורכבות העימוד שלהם מה-MSA האמתי, כלומר, המורכבות של MSAs מסומלצים היו נמוכים בהרבה מזו של האמתי. שלישית, התגלה כי להכללה של רצפים קצרים, חלקיים, יש השפעה משמעותית על מורכבות העימוד. רביעית, מצאתי כי התפלגות השpower-law המשמשת בדרך כלל לתיאור ההתפלגות של אורכי indels, אינה בהכרח המתאימה ביותר לכך. לסיכום, המסקנה שלי היא כי התחום של סימולציית רצפים נמצא בחיתוליו, ומחקר רב עדיין נדרש כדי להשיג סימולציות מציאותיות יותר. עד אז, המסקנות המתוארות על בסיס הסימולציות הנוכחיות צריכות להילקח בזהירות רבה.