https://doi.org/10.35662/unine-thesis-3096

Frequentist Estimation of Evolutionary History of Sequences with Substitutions & Indels

A dissertation submitted to

UNIVERSITY OF NEUCHÂTEL

In partial satisfaction of the requirements for the degree of

DOCTOR OF SCIENCE

Presented by

Gholamhossein Jowkar

Committee in charge

Prof. Dr. Daniel Croll, University of Neuchâtel, Switzerland (Co-chair)

Prof. Dr. Maria Anisimova, Zürich University of Applied Sciences, Switzerland (Co-chair)

Prof. Dr. Pilar Eugenia Junier, University of Neuchâtel, Switzerland (Internal expert)

Prof. Dr. Ziheng Yang, University College London, UK (External expert)

19.03.2024

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Gholam-Hossein JOWKAR

Titre :

"Frequentist Estimation of Evolutionary History of Sequences with Substitutions & Indels"

sur le rapport des membres du jury composé comme suit :

- Prof. Daniel Croll, co-directeur de thèse, Université de Neuchâtel, Suisse
- Prof. Maria Anisimova, co-directrice de thèse, ZHAW, Wädenswil, Suisse
- Prof. Pilar Junier, Université de Neuchâtel, Suisse
- Prof. Ziheng Yang, UCL, United Kingdom

Neuchâtel, le 4 mars 2024

Le Doyen, Prof. R. Bshary

Abstract

Estimation of the evolutionary history of molecules is mainly done by reconstructing the ancestral sequences given present-day sequences and phylogeny information. Biological sequence data is a result of evolution by mutational events such as character substitutions (or point mutations), insertions and deletions (indels). Inference of the evolutionary history of sequences with substitution and indels can be used in various biomedical applications, from tracking the origin of pandemic viruses to studies of the cause of visual impairment.

Indels are among the most important sources of genomic variation and carry sound evolutionary signals; however, well-known ancestral sequence reconstruction (ASR) methods ignore or mistreat them. ASR with indels is a big challenge from both computational and statistical viewpoints. This research proposed a novel solution to infer the ancestral sequences, while accounting for the evolutionary indel process.

First, I used an evolutionary model of substitution and indel for ASR and implemented it in the ARPIP program. ARPIP implemented a novel empirical Bayes method, which allows us to reconstruct ancestral sequences with indels under the Poisson indel process (PIP). While PIP is a continuous-time Markov chain (CTMC) model that assumes single-character indels, and has important computational advantages. I showed that ARPIP reconstructed biologically reasonable indels.

Second, it is difficult to model multiple-character (or "long") indels since most evolutionary CTMC models assume site-independence. Thus, I investigated whether a single-character indel assumption was detrimental for ASR. Analysis of real and simulated data showed that the single-character indel model could be used for ASR. ARPIP preserved gap length distribution in multiple sequence alignment, including regions with long indels. Moreover, the indel variation in six *eutherian* mammalian orthologous proteins was studied to explore the evolutionary dynamics of insertions and deletions.

Finally, ASR, similar to other inferences, is affected by uncertainty. To account for it, a posterior probability profile method was devised. In collaboration with an experimental lab to study properties of ancestral proteins, the approach was applied to reflect the variation in ASR inference on *neural retina leucine zipper* transcription factor of selected vertebrates. Moreover, an alternative reconstruction for the ambiguous regions was introduced.

Keywords: deletion, insertion, joint ancestral sequence reconstruction, Poisson indel process, uncertainty of inferred ancestral sequences.

Résumé

L'estimation de l'histoire évolutive des molécules se fait principalement en reconstruisant les séquences ancestrales à partir des séquences actuelles et des informations phylogénétiques. Les données sur les séquences biologiques sont le résultat de l'évolution par des événements de mutation tels que les substitution, les insertions et les délétions (indels). L'inférence de l'histoire évolutive des séquences avec substitution et indels peut être utilisée dans diverses applications biomédicales, de la recherche de l'origine des virus pandémiques à l'étude de la cause des déficiences visuelles.

Les indels font partie des sources les plus importantes de variation génomique et sont porteurs de signaux évolutifs importants ; cependant, les méthodes bien connues de reconstruction des séquences ancestrales (ASR) les ignorent ou les traitent de manière incorrecte. La reconstruction des séquences ancestrales avec les indels est un grand défi, tant du point de vue informatique que du point de vue statistique. Ma recherche propose une nouvelle solution pour déduire les séquences ancestrales tout en tenant compte du processus évolutif des indels.

Tout d'abord, j'ai utilisé un modèle évolutif de substitution et d'indel pour l'ASR et je l'ai implémenté dans le programme ARPIP. ARPIP a mis en œuvre une nouvelle méthode empirique de Bayes, qui nous permet de reconstruire les séquences ancestrales avec des indels dans le cadre du processus de Poisson (PIP). Le PIP est un modèle de chaîne de Markov à temps continu (CTMC) qui suppose des indels d'un seul caractère et présente d'importants avantages en termes de calcul. J'ai montré que l'ARPIP reconstruisait des indels biologiquement raisonnables.

Deuxièmement, il est difficile de modéliser des indels à caractères multiples (ou "longs") car la plupart des modèles évolutifs CTMC supposent l'indépendance des sites. J'ai donc cherché à savoir si l'hypothèse d'un indel d'un seul caractère était préjudiciable à la ASR. L'analyse de données réelles et simulées a montré que le modèle indel à caractère unique pouvait être utilisé pour la ASR. L'ARPIP a préservé la distribution de la longueur des écarts dans l'alignement de multiples séquences, y compris dans les régions comportant de longs indels. En outre, la variation en indels de six protéines orthologues chez des mammifères *eutherian* a été étudiée pour explorer la dynamique évolutive des insertions et des suppressions.

Enfin, l'ASR, comme d'autres inférences, est affectée par l'incertitude. Pour en tenir compte, une méthode de profil de probabilité a posteriori a été mise au point. En collaboration avec un laboratoire expérimental qui étudie les protéines ancestrales, l'approche a été appliquée pour refléter la variation de l'inférence ASR sur le facteur de transcription *neural retina leucine zipper* de vertébrés sélectionnés. En outre, une reconstruction alternative pour les régions ambiguës a été introduite.

Mots-clés: délétion, insertion, reconstruction conjointe de séquences ancestrales, processus d'indel de Poisson, incertitude des séquences ancestrales déduites.

Acknowledgements

I would like to express my sincere thanks and appreciation to the following for their advice, support, and invaluable guidance during the completion of the present thesis.

First of all, I am grateful and thankful to Almighty God who helped me to finish this thesis.

I am very much thankful to Professor Maria Anisimova, my supervisor, for her believing in me since the beginning of the project and also for introducing me to the field of computational molecular evolutionary biology. Her effective guidance, constructive criticisms, and constant encouragement throughout this thesis are also highly appreciated.

Many great thanks are also due to Dr. Manuel Gil, my advisor, for his worthwhile guidance and support in accomplishing this research.

I am also grateful to Professor Daniel Croll for helping and supporting me during my studies.

I am also very pleased to thank Professor Ziheng Yang and Professor Pilar Junier for agreeing to join the committee for my PhD examination.

I am thankful to Dr. Massimo Maiolo and Dr. Jūlija Pečerska for their valuable and helpful counseling in the process of working on the research.

I would like to thank my friend, Erich Zbinden and his wife Catherine Zbinden, for acquainting me with the Swiss culture. Also, sincere thanks to Dr. Markus Bott and Agnieszka Wronska, my good friends, who were always encouraging and kind to me.

I would like really to thank my parents, who encouraged and supported me to start, continue, and finish my studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Estimating evolutionary history at the molecular level is to infer the past state of molecular sequences given present-day information and phylogenetic models. This estimation has a wide range of applications, from biomedicine [Zakas et al. 2017] to protein engineering [Thomson et al. 2022]. Most importantly, studying molecular evolutionary history may help us understand gene functions and discover how various genetic traits evolve, resulting in an understanding of different mutation mechanisms, which contribute to species divergence. Reconstructing the molecular history could facilitate studies of medically relevant genes in humans and other species. Restoring the molecular pathways during evolution could help to investigate disease-related genes. This type of study also provides the means to design new treatment strategies and drugs for genetic diseases.

With the rapid expansion of genomic data, statistical and computational methods play a crucial role in inferring the evolutionary history of sequences with mutations. Markov-based evolutionary models provide the means to describe sequence changes (substitution and indel mutations). Many algorithms and methods were implemented based on these evolutionary models in a probabilistic setting.

The mechanisms of insertion and deletion (indel) mutation have been studied less than those associated with substitution mutations [Kamneva et al. 2010]. This has been exacerbated by the fact that most existing methods ignore, penalize, or treat indels as missing or ambiguous data. Indels, however, are known to have much of a phylogenetic signal and should be important in phylogenetic applications, including when reconstructing molecular histories of genomic sequences. Therefore, the central challenge of this estimation is to model indels as biologically and phylogenetically reasonable as possible.

To my knowledge, Poisson Indel Process (PIP) [Bouchard-Côté and Jordan 2013] serves this purpose well by modeling indels biologically, phylogenetically, and statistically sensible. From an evolutionary point of view, this model has a biological and phylogenetical interpretation, as the indel process is defined as a stochastic process on a phylogeny, in addition to substitutions. From the computational point of view, the PIP computes the joint likelihood of the tree and MSA in linear time with respect to the number of taxa. This evolutionary model provides us with the means to model the evolutionary history of sequence with substitution and indels and will be used in this thesis for all presented new methods and analyses.

Ancestral sequence reconstruction (ASR) allows us to estimate the evolutionary history from a set of biological sequences, but with many challenges. A major challenge in ASR is to model indels with various lengths adequately, which makes it a harder problem in comparison with substitutions. Another central challenge of phylogenetic tasks that exist in nearly all the inferences, including ASR, is uncertainty. These are the main challenges this thesis tries to address.

This thesis deals with both the computational method developement and applications of this new methodology to real data. Foremost, I present a maximum likelihood (ML) joint ancestral sequences reconstruction (ASR) with substitution and indels under PIP. Next, I introduce an exploratory analysis of indel patterns in six mammalian orthologous proteins to show the patterns of indels in mammalian protein evolution. This study demonstrates that PIP, as a single-character evolutionary model, can nevertheless describe multiple-character indels in ancestral sequences. Eventually, I propose how to measure the uncertainty in ancestral sequences inference by introducing probability profiles that reflect the confidence in the inferred ancestral state.

1.1 Statement of the problem

Phylogenetic inference can be divided into several interrelated tasks performed as independent steps, including multiple sequence alignment (MSA), phylogeny tree estimation, detection of selection, substitution model estimation, and ASR. ASR methods infer the ancient sequences given an evolution model of mutations and contemporary sequences and phylogenetic tree [Joy et al. 2016]. We assume that our input (i.e., MSA, tree topology, branch lengths, substitution model, transition probability as a function of substitution model, and branches of the tree) are accurate. However, it is known that all these inputs could be potential sources of error (for e.g., see [Cunningham et al. 1998] studeid potential sources of error in the branch lengths estimation).

Up to now, evolutionary patterns generated by indel events have been studied less than sub-

stitution [Kamneva et al. 2010]. However, indel mutations are abundant in molecular biology datasets and carry important evolutionary signals [Tao et al. 2007]. Indels have functional consequences such as decreasing transcription and abnormal protein aggregation [Sehn 2015]. Moreover, indel processes are a rich source of genetic variation in different species, including the human genome [Mills et al. 2006, Sehn 2015].

A main limitation of all the existing ASR methods is handling indels adequately in ancestral lineages, which requires an accurate and more realistic model of the evolution of the sequences with insertion and deletion events. Most statistical ASR methods ignore indels or treat them as missing data by removing the columns containing gap character [Yang 1997, Pupko et al. 2000]. This is another potential source of error in current ASR models. Those that treat indels typically provide an unrealistic evolutionary model of indels [Kishino et al. 1990, Ashkenazy et al. 2012, Holmes 2017]. Computational complexity is always an issue when implementing evolutionary models. Moreover, there is a trade-off between the more realistic model of indels and the complexity of the model. The existing indel models deployed are either computationally time inefficient (e.g., [Diallo et al. 2009]) or biologically unrealistic (e.g., [Ashkenazy et al. 2012]).

The existing evolutionary models of substitutions are based on the continuous-time Markov-chain (CTMC) theory, assuming site-independence. Compared to substitutions, multiple-character indel violates the site-independence assumption as they tend to appear in a sequence alignment as long fragments of gaps. Therefore, modeling multi-character (long) indels is particularly challenging. As has been mentioned, there is a trade-off between the realistic model of indels and the complexity of the model. The more realistic the models (without site-independence assumptions), the more computationally complex and impossible to deploy for large-scale datasets (e.g., [Thorne et al. 1992]). Notice that models such as TKF92 are not fully biologically realistic, as there is no possible explanation for overlapping indel events. However, computing phylogenetic likelihoods under TKF92 has exponential complexity.

Most ASR tools only aim to infer the most probable ancestral (MPA) character state (with few exceptions, e.g., [Yang et al. 1995]). The ASR methods generally select the state that maximizes or minimizes the given criteria (in a probabilistic framework, it is the state with the highest posterior probability). Selecting the optimal MPA while ignoring the sub-optimal MPAs is a well-known source of bias in ASR [Yang 2014]. Indeed, multiple characters can have the highest probability with slight differences due to numerical computation or cumulative errors. Ignoring this variability in reconstructed ancestral states will be considered a limitation of the current inference techniques. Ancestral sequences are usually inferred, given a fixed MSA and a phylogenetic tree. Separate inferences of MSA, tree and ancestral states can lead to the accumulation of errors and biases in the inference. One possible reason for this is the lack of an explicit model of indel in the stages of MSA and tree estimation. Substantial uncertainty

is often associated with inferences of alignment and phylogeny.

Insertions and deletions are typically difficult to disentangle in genomic sequences, so they are poorly treated in the phylogenetic methods. The lack of proper modeling of indel events hinders all types of phylogenetic analyses. The evolutionary dynamics of indels are highly informative as indels are commonly implicated in constitutional (hereditary) and somatic (acquired, including cancer) diseases [Sehn 2015]. Inferring substitution and indel histories over time enables studies of the evolution and function of ancient molecules [Pagel 1999, Thornton 2004]. Indel analysis has applications in diagnosing genetic diseases, patient counseling, therapy selection, and prediction of patient prognosis [Sehn 2015].

Ignoring uncertainty in the inference is a serious limitation of current ASR techniques [Oliva et al. 2019]. Conditioning the ASR on inaccurate MSA or tree will introduce biases to the inference. Even when using an ASR method with an adequate model of indels, these biases in estimation parameters can result in overconfident MPA sequences. The potential uncertainty in the inference is not reflected in the MPA, even though it is convenient to get just a point estimate. Using biased or inaccurate MPA for experimental research is problematic due to the significant expense of synthesizing and expressing such sequences.

This dissertation proposes a solution to the problem of estimating the evolutionary history of sequences that is more realistic and more accurate. This research's objective is to reconstruct the ancestral sequence along a given phylogenetic tree and sequence alignment using models that include both substitutions and indels. The thesis uses the PIP model to describe insertions and deletions and proposes a way of using it for ASR. The aim is to reduce the errors due to the mistreatment of the indels using PIP while reflecting the uncertainty in the MPA through the probability profile.

1.2 Contributions

The main contribution of this thesis to computational molecular evolutionary biology research can be divided into three parts: A new method for ASR with indels, a simulation study and a large-scale analysis of indel dynamics in real data, and a new approach to treating uncertainty of ASR inference.

- Chapter 3 (**New ARS method**): Most of the existing ASR tools treat gaps as missing/ambiguous data or remove gap characters entirely. Ignoring them would result in the loss of evolutionary information. Some other methods tried to model indels, which were not biologically interpretable, phylogenetically reasonable, or computationally tractable. This new algorithm overcomes all three using the PIP to model substitutions and indels

in linear time. To infer ancestral sequences with indels, this thesis provides a tool that: 1) uses the PIP model for joint reconstruction of ancestral character states, including indels. 2) is implemented in the ML framework, i.e., we use an empirical Bayesian approach with ML estimates. 3) validate the method by simulations and demonstrate its performance in simulations and on a real-world SARS-CoV dataset.

- Chapter 4 (**A simulation study**): This chapter investigates the pattern of indel evolution in simulated data and in mammalian orthologs. This thesis provides evidence that the single-character indel model PIP preserves the gap length distribution. This exploratory analysis has been performed on six *eutherian* confirmed the well-established deletion bias.

- Chapter 5 (**Measuring uncertainty of ASR inference**): Existing statistical ASR methods typically only provide the MPA. Although accurate MPA reconstruction is the goal of ASR from an inferential perspective, given the high costs of synthesizing and expressing in vitro, the accuracy of ASR methods is still insufficient, particularly for more divergent sequences. The biological expert can measure the uncertainty of reconstruction using a probability profile. This would provide a chance to consider the variability of ancestral inference rather than using a single MPA with error. The computational approach is demonstrated on 81 species of vertebrates *neural retina leucine zipper* (NRL) transcription factors (TF). Using ARPIP, I have inferred the ancestral sequence of selected vertebrates with their confidence probability profiles.

1.3 Scientific outputs

The following section lists the scientific contributions as part of the research presented in this thesis.

1.3.1 Manuscripts

- "A Probabilistic Solution to Measure the Uncertainty of the Ancestral Sequences for Squamates Neural Retina Leucine Zipper Transcription Factor". **Gholamhossein Jowkar**, Manuel Gil, and Maria Anisimova. Technical report, 2023.

- "Single-character insertion-deletion model preserves long indels in ancestral sequence reconstruction". **Gholamhossein Jowkar**,Julija Pečerska, Manuel Gil, and Maria Anisimova. The manuscript draft was uploaded to bioRxiv and submitted to a peer-reviewed journal. DOI:10.1101/2024.03.09.584071

- "ARPIP: Ancestral sequence Reconstruction with insertions and deletions under Poisson indel process". **Gholamhossein Jowkar**, Julija Pečerska, Massimo Maiolo, Manuel Gil, and Maria Anisimova. Systematic Biology, 2023. DOI:10.1093/sysbio/syac050

1.3.2 Talks and posters

Below is the list of presented research talks and posters at scientific conferences and events:

- Phylogenetic Ancestral Reconstruction with indels - February 2020 - ZHAW, Switzerland. (Talk)
- Poisson Indel Process for modeling indels in molecular sequences– March 2020 – University of Neuchatel, Switzerland. (Talk)
- Phylogenetic Ancestral Reconstruction Under Poisson Indel Process, SIBdays 2020 – 19 June 2020 – online, Switzerland. (Poster)
- Phylogenetic Ancestral States Reconstruction Under Poisson Indel Process - September 2020- University of Neuchatel, Switzerland. (Talk)
- Phylogenetic Ancestral Reconstruction with ARPIP - April 2020 - ZHAW, Switzerland. (Talk)
- Estimation of Evolutionary History of Sequences with Substitutions and Indels- September 2021 – ZHAW, Switzerland. (Talk)
- ARPIP: Ancestral sequence Reconstruction with insertions and deletions under the Poisson Indel Process, MCEB (Mathematical and Computational Evolutionary Biology) Conference – 26-30 June 2022 – Château d'Oex, Switzerland. (Poster)
- Patterns of Insertion and Deletion Events in Six Mammalian Orthologous Proteomes using Poisson Indel Process – September 2022 – Au, Zurich, Switzerland. (Talk)
- Patterns of Insertion and Deletion history under Poisson Indel Process - SIB PhD annual retreat, October 2022- Fribourg, Switzerland. (Talk)
- Patterns of Indel Events in Six Mammalian Orthologs using Poisson Indel Process, SMBE Conference- 23-27 July 2023- Ferrara, Italy. (Poster)

1.3.3 Software packages

The two software packages developed during this research are as follows:

1.3.3.1 ARPIP: Ancestral sequence Reconstruction under Poisson Indel Process

ARPIP is a method for ASR that uses the PIP [Bouchard-Côté and Jordan 2013] to model single-site insertions and deletions on a phylogenetic tree, assuming independence among sites. In addition to the MPA, ARPIP outputs the probability profile corresponding to each inferred ancestral state to tackle reconstruction uncertainty. The proposed algorithm has been implemented based on Bio++ open source library [Guéguen et al. 2013] using C++ programming language. The code is freely available at `https://github.com/acg-team/bpp-ARPIP` under the GNU GPLv3 license.

1.3.3.2 Indelviewer: Visualizing reconstructed indel events

IndelViewer is a Python library to facilitate visualizing indel events for ARPIP outputs which could help researchers to better understand the evolutionary process in targeted case studies. This package uses the ETE3 python library [Huerta-Cepas et al. 2016] and Bokeh [Bokeh Development Team 2018] to visualize the indel event on the tree structure. IndelViewer is an interactive stand-alone or web app to scroll through MSA columns and see the corresponding indel events on the tree topology. The Indelviewer code is distributed under General Public License and can be obtained from `https://github.com/acg-team/IndelViewer`.

1.4 Organization of the thesis

This dissertation is organized as follows:

Chapter 2 provides an overview of the necessary prerequisites and theoretical background for understanding this thesis. This chapter begins with preliminary content, including multiple sequence alignment, phylogenetic tree, substitution, insertions, and deletions. Then, it continues with evolutionary models and statistical methods to compute the evolutionary history of sequences.

Chapter 3 has been published in the "Systematic Biology" journal. This work introduces a novel ASR algorithm with substitution and indels. This algorithm and its open-source software are the main contributions of this thesis.

Chapter 4 includes a manuscript draft that is submitted for peer review. This chapter examines gaps and indel patterns in orthologs from six mammalian species and in simulated data. This

part of the research studied modeling long indel using PIP, a single-character model.

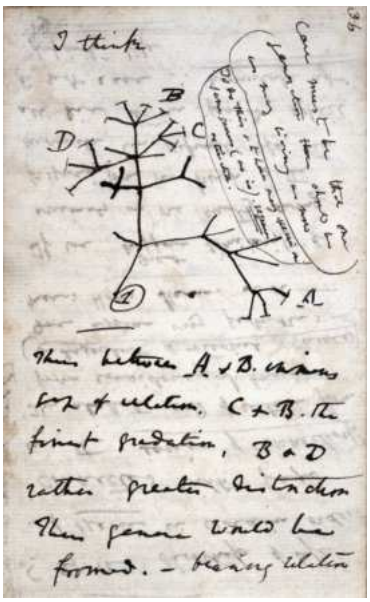
Chapter 5 presents a probabilistic analysis of the ancestral sequence of Squamates, addressing questions based on our collaboration with the experimental biology team of Prof. Belinda Chang (University of Toronto). This chapter examines reconstructed ancestral sequences of a group of Squamates's retina protein. This work introduces the probability profiles to reflect the reconstruction uncertainty in ancestral character states.

Chapter 6 summarises the thesis findings and discusses a guideline for future research opportunities.

Chapter 2

Background

In this chapter, a comprehensive overview of the theoretical and methodological aspects of this thesis will be provided. The concept of evolution of species, multiple sequence alignment, phylogenetic trees and substitution models will be introduced. Lastly, the ASR task with its mathematical and computational elements, including the treatment of gaps, will be presented.



From Darwin's Notebook now stored in Cambridge University library

2.1 Evolution and homology

Charles Darwin's theory of evolution is one of the most prominent theories in the field of biology. Darwin, in his book entitled "The Origin of Species", introduced the mechanism of natural selection [Ruse 1975]. Natural selection postulates that some organisms, based on their useful characteristics, have a higher chance of survival and reproduction. Consequently, these useful traits will get passed more often than other traits [Ruse 1975]. Darwin showed how natural selection leads to evolutionary changes through time. Nowadays, the theory of evolution is being used to solve problems in medicine, geology, paleontology, and demography.

Evolutionary thinking provides us with the tools to analyze and discover the molecular mechanisms of species divergence. Since the introduction of Darwin's theory of evolution, the evolutionary relationship has been typically described by a tree structure called phylogeny. Usually, similar protein sequences or genes have a common evolutionary origin. Phylogeny normally represents the evolution of protein sequences or genes with a common origin or ancestor. Sequences with shared ancestry are called homologs.

Molecular evolutionary studies typically start with identifying homologous regions within or across species. We define homology as a hypothesis, based on observed sequence similarity, where nucleotide or protein sequences are similar because they have a common evolutionary origin [Altenhoff and Dessimoz 2012]. Characters from a common ancestor are called homologous characters. Homologous characters may differ because of mutations [Bouchard-Côté 2010].

Two powerful forces shaping the evolutionary divergence of genes with shared ancestry are speciation and gene duplication [Jensen 2001], which would produce different homologous evolutionary relationships: 'ortholog', and 'paralog' (Fig. 2.1).

Orthologs are pairs of genes that originated by speciation (see Fig. 2.1). In general, orthologs are believed to retain the same function during evolution. Thus, identifying orthologs is a critical process for reliable gene function prediction in newly sequenced genomes [Jensen 2001].

Paralogs are pairs of genes that started diverging via gene duplication and often belong to the same species, but this is not always the case. Paralogs, unlike orthologs, are believed to have slightly different functions. Concerning the application, orthologous genes are often used to infer function or species trees, while paralogous genes are commonly used to study function innovation [Altenhoff et al. 2019a].

Lastly, analogs are pairs of non-homologous genes with identical or similar functions. Analogs do not share a common ancestor (see Fig. 2.1).

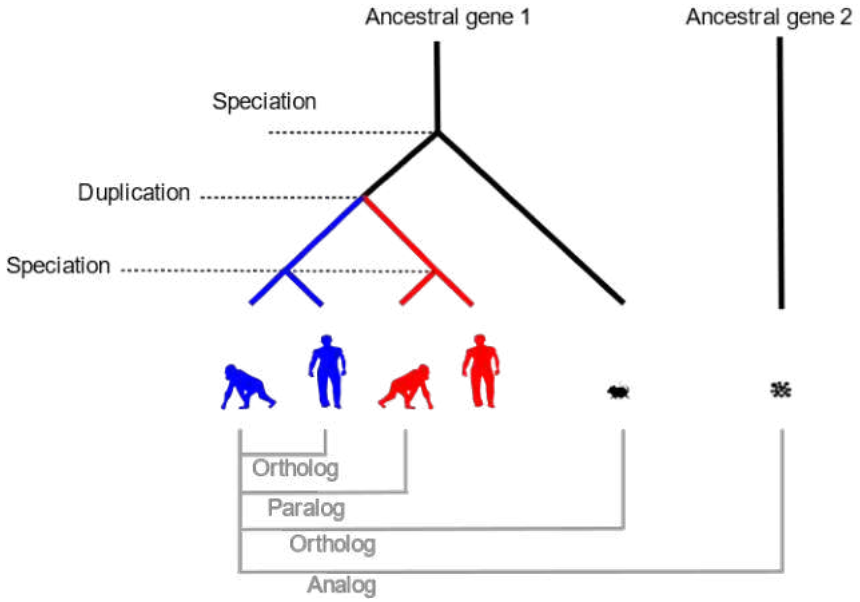


Figure 2.1: A tree of homologous genes, with speciation and gene duplication events. Genes that diverged by speciation are called orthologs, while genes that diverged by duplication are called paralogs. Analogous genes are similar but do not share common ancestry.

2.2 Species, sequences, and multiple sequence alignment

A biological species is a group of organisms that can interbreed. They form natural taxonomic units.

From a biological point of view, phenotypic traits are derived from the genetic information inherited from the ancestor. Genomic sequences represent the genetic characteristics of species. The sequences carry genetic information and pass it to the offspring. Such sequences can be formed from nucleotides (DNA or RNA), amino acids (AA), or codons. Each level of abstraction is represented by an alphabet. There are4 nucleotides. Triples of nucleotides are called codons. There are 64 such triples. In the universal genetic code, 61 triples code for the 20 AAs (in a redundant way), whereas 3 triples code for stop codons. Stop codons are typically not considered in phylogenetic analysis. We denote such an alphabet by Σ.

Almost all sequence analyses start with alignment, i.e., the identification of homologous characters in homologous sequences. This is necessary due to the insertion-deletion process (reviewed below). A multiple sequence alignment (MSA) of a set of three or more homologs consists of a matrix M. M over the alphabet Σ and plus the gap symbol '-'. The n rows represent genomic sequences (for example, representing different species), and the m columns are the homologous characters (see Fig 2.2). In this thesis, the MSA is assumed to be given while I focus on the ASR task.

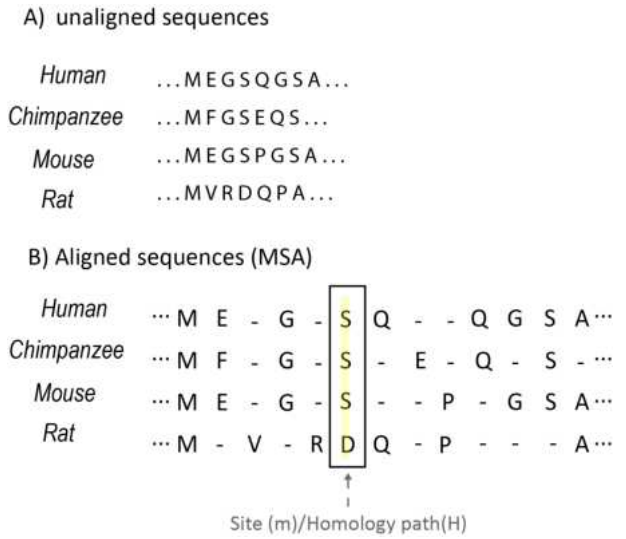


Figure 2.2: Unaligned sequences (A) from four species and a corresponding MSA (B)

2.3 Phylogenetic tree

A phylogenetic tree, or phylogeny, is a tree diagram describing common ancestry. From the mathematical point of view, a tree $\tau = (\mathscr{V}, \mathscr{E})$ is a connected acyclic graph consisting of a set vertices (or nodes)$\mathscr{V}$ and a set of edges (or branches) $\mathscr{E}$. Each edge connects a pair of vertices ($\mathscr{V} \times \mathscr{V}$), and there is exactly one path between every two vertices. Each internal node in a tree represents a speciation or duplication event corresponding to the most recent common ancestor (MRCA) of all taxa under that node. For each $v \in \mathscr{V}$, the branch length $b(\mathrm{v})$ is a positive real number proportional to the time that separates two events (see Fig. 2.3).

Phylogenetic trees can be either unrooted or rooted. Unrooted trees make no assumption on ancestry but only evolutionary relatedness. Rooted trees, on the other hand, include ancestry.

The root in the tree is the node with the degree of two denoted by Ω. A rooted phylogenetic tree uses the MRCA of all taxa to provide the direction of the evolution process. In this representation, $\Omega \rightarrow v$, v is called the descendant, and Ω is called the ancestor of node v. In a rooted bifurcating tree, all internal nodes have exactly two descendants and one ancestor, except for the root, which has two descendants but no ancestor. The set of leaves $\mathscr{L}$ are the vertices of degree one in the tree. Given $|l|$ leaves ($l \in \mathscr{L}$) $l-1$ internal nodes in a rooted tree τ (including the root).

The branch lengths typically represent the expected number of substitutions per site. The clock assumption of trees says that we have a constant rate of evolution across branches [Yang and Rannala 2012]. The sum of branch lengths along a path from the root to the leaves under the non-clock assumption is not required to be constant [Bouchard-Côté 2010]. In this research,

the phylogeny is assumed to be given with the condition that all the trees are bifurcated rooted trees with non-clock assumptions.

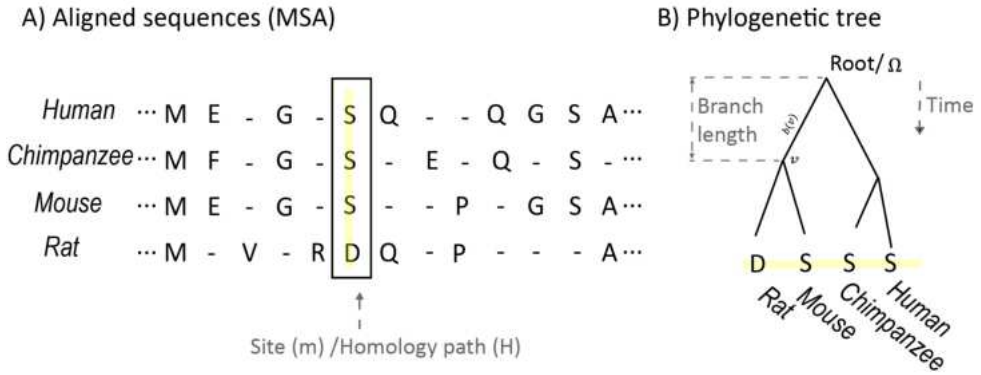


Figure 2.3: Toy example of an MSA for a set of sequences and the phylogenetic relationship of a site m.

2.4 Mutation

Biological researchers try to understand the gene's encoded instructions by identifying genetic variation between genomes of species or individuals. One fundamental mechanism that increases genetic variation is mutation [Tao et al. 2007]. Mutation in molecular biology is the process of changing the nucleotide sequence, which results in new and heritable characteristics. Evolutionary biologists reconstruct the history of these changes to understand how their function changes over time [Macdonald et al. 2022].

Mutations have a variety of effects on downstream protein production. These effects might be neutral, beneficial, deleterious, or lethal. Many mutations are neutral as they have neither negative nor positive effects on the organism in which they occur. Beneficial mutations are essential for evolution to occur.

Mutation sometimes leads to new versions of proteins that help organisms adapt to environmental changes. They increase an organism's chances of surviving or reproducing, so they will likely become more common over time due to positive selection. Mutations can also prevent one or more proteins from working correctly. By changing a gene's instructions for making a protein, a variant can cause a protein to malfunction or not to be produced at all, causing genetic disorders.

Concerning the gene structure, mutations can be either small or large scale. Large-scale mutations (including deletions, duplications, inversions, insertions, and translocations) affect large chromosome regions, whereas, in small-scale mutations, one or a few nucleotides are affected.

2.4.1 Small-scale mutations

The commonly used model of small-scale mutations includes substitution, insertion, and deletion.

2.4.1.1 Substitution

Point mutations are replacements of a single character, either a single base or amino acid (AA). A point mutation could become fixed in a population, either by chance or through the process of natural selection. Point mutations that get fixed in populations are typically referred to as substitutions.

At the DNA level, two categories of this variation of mutation are transition and transversion. The former refers to substituting a purine for another purine ($A \rightarrow G$) or one pyrimidine for another pyrimidine ($C \rightarrow T$). The latter refers to replacing a purine with a pyrimidine or vice versa (for example, $C \rightarrow G$). Point mutations usually occur during DNA replication. Such a mutation might change the encoded AA and the downstream protein production (see Fig. 2.4. B).

2.4.1.2 Insertions and deletions

Insertions refer to a type of mutation characterized by inserting a sub-sequence of DNA into a DNA sequence. Deletions delete a sub-sequence of DNA from a sequence. Insertions and deletions (indels) are represented by gaps in sequences, ranging from 1 to $1K$ base pairs in length [Sehn 2015]. Indels often happen in chromosome regions due to the DNA polymerase slipping during DNA replication. Indels affecting protein-coding sequences could have more severe effects as they can cause frame-shifts. For example, an in-frame mutation occurs when indels happen in groups of three bases while preserving the reading frame. The frame-shift mutation changes the reading frame used to translate the sequence into a protein product, therefore changing the resulting protein. The earlier the indel mutation occurs in the sequence, the more the protein is altered in downstream production (see Fig. 2.4 C and D). This means that while the reading frame remains intact, depending on the size of the indel, multiple new AAs might affect the protein's function. For example, out-of-frame mutations can create a premature or postmature stop codon and a shorter or longer polypeptide.

2.4.2 Large-scale mutations

Large-scale mutations involve a higher level of genetic material change in the chromosome. Well-known large-scale mutations would be classified as gene duplications, deletion of large chromosomal regions, and chromosomal inversions. Gene duplications are mutations where

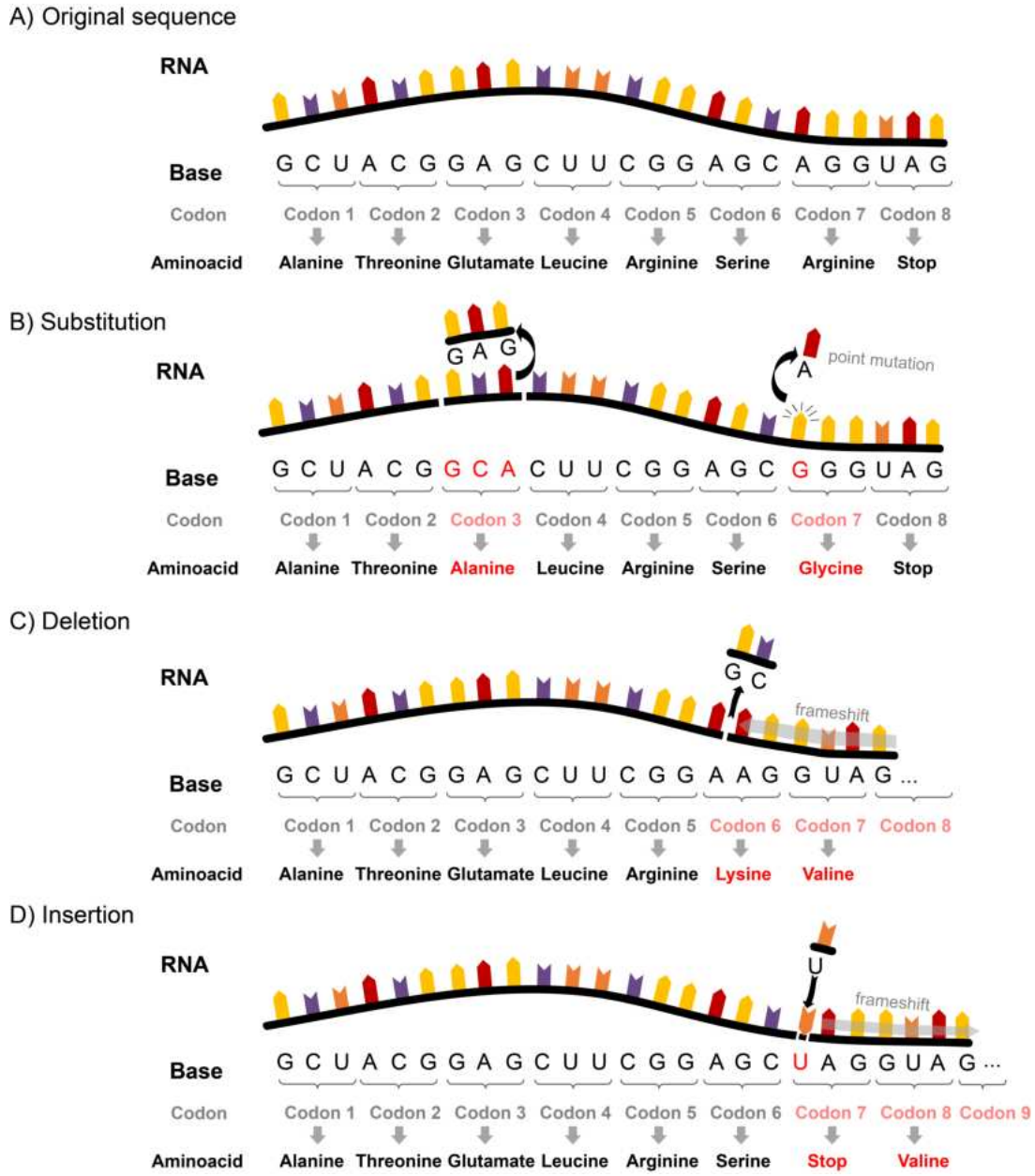


Figure 2.4: Depiction of different small-scale mutations: A) Original sequence B) Substitution at the nucleotide and AA level. C, D) Indel at two sequences by multiple nucleotide deletions (C) or single nucleotide insertion (D). Due to the indel event, we can see mutation in the sequence completely altered the base sequence, resulting in the formation of completely different AAs during the translation process.

an extra copy of a chromosome region or gene is generated. Chromosomal inversions are mutations where a rearrangement within the chromosome causes a segment to be end-to-end reversed. There are other known large-scale mutations, such as chromosomal translocation and copy number variation.

Mutations are essential for evolution because they increase genetic variation and the potential for individuals to differ. In this thesis, we only consider substitutions and indels mutations.

2.5 Mutation model used in this thesis

This section briefly reviews the mutation model considered in this thesis. It is a stochastic process where the characters (i.e., Σ) change along the phylogenetic tree τ in the direction of the root (Ω) to the leaves (L).

Typically, evolutionary models consider substitution and indels. This section introduces continuous-time Markov chains (CTMC) used to model substitutions and then explores statistical approaches to extend this model to contain indels. For the remainder of this thesis, we will assume that the evolutionary process is applied to a rooted bifurcation phylogenetic tree.

2.5.1 Markov model of substitution

This section introduces the continuous-time Markov chain (CTMC) model basics, while more information can be obtained in [Tavaré 1986]. A CTMC is a probability model of changes (substitution) of alphabet Σ of size n and is defined by $n \times n$ transition-rate matrix (or generator matrix) Q. The transition probability matrix P over the time $t \geq 0$ from state i to state j is computed by:

$$P(t) = \exp(Qt), \tag{2.1}$$

Where the generator matrix Q is a symmetric matrix measuring substitution rates:

$$Q_{ij} = \begin{cases} q_{ij}, & \text{if } i \neq j \\ -\sum_{k \neq i} q_{ik}, & \text{if } i = j, \end{cases} \tag{2.2}$$

Moreover, all q_{ij} in the transition-rate matrix Q should be non-negative ($q_{ij} \geq 0$; for $i \neq j$ and $1 \leq i, j \leq |\Sigma|$) and the diagonal entries ($0 \leq -q_{ii} < \infty$) are set to ensure that each row of $P(t)$ is well defined as a probability mass function ($\sum_j q_{ij} = 0$; for all i).

The Markov memoryless property says that the transition probability to the next state depends only on the current state and is independent of previous states. The process is stationary and time-reversible, and for computational convenience, the independence of sites is assumed. Its stationary distribution (background frequency) π satisfy the following equation:

$$\pi Q = 0, \tag{2.3}$$

where $\sum_{i\in\Sigma} \pi_i = 1$. This process is reversible, i.e. for all $1 \leq i, j \leq |\Sigma|$:

$$\pi_i q_{ij} = \pi_j q_{ji}, \tag{2.4}$$

The CTMC models considered in this thesis are time-reversible, so the direction of the substitutions cannot generally be identified from the data. Although this thesis uses rooted phylogenies, we still benefit from the CTMC's stationarity and reversibility.

2.5.2 Evolutionary model of substitutions and indels

Historically, the existing evolutionary models assume that the indels are independent of the substitutions. However, it is questionable whether this assumption is correct, as some models infringe it [Bouchard-Côté and Jordan 2013, Thorne et al. 1991].

Thorne-Kishino-Felsenstein 1991 (TKF91) [Thorne et al. 1991] is a pioneer CTMC evolutionary model of indels and substitution; however, its proposed solution is computationally expensive. The PIP [Bouchard-Côté and Jordan 2013] was later introduced as a simplification of the TKF91 and computationally efficient. The abstract concept of these two molecular evolutionary models would be as follows:

2.5.2.1 TKF91 model

TKF91 is the first evolutionary sequence model that allowed indels under a simple birth-death process [Thorne et al. 1991]. This model allows for single-site indels only. Despite this, TKF91 is too costly and, therefore, not applicable to large datasets due to the exponential time of marginal likelihood computation.

In TKF91, the sequence alphabet is considered to be pairs of alphabet letters (characters) attached by immortal links to the leftmost character. The immortal link (cannot be deleted) is associated with the birth process with rate λ_{TKF91}, whereas mortal links (can be deleted) are associated with both the birth process of rate λ_{TKF91} and death process with rate μ_{TKF91}

($\lambda_{\text{TKF91}} > \mu_{\text{TKF91}}$ prevents vanishing sequence over the evolution time). The presence of the immortal link is necessary to prevent a vanishing sequence or having a sequence with an infinite length.

Given a pair of sequences, TKF91 allows computing transition probability by initiating two subprocesses: substitution and indel. To simulate the mutation processes (substitution, insertion, and deletion) on the sequence of length N, we assume $3N+1$ independent random variables with exponential distribution [Bouchard-Côté 2010].

The substitution process is a modified reversible CTMC substitution model. The N random variables are the transition probability for normal links due to substitution processes (see [Thorne et al. 1991] for more explanation).

The indel process is defined on the links between characters rather than the characters themselves. The birth-death process applies to each link independently and increases or decreases sequence length by one unit.

The indel process assumes independence among sites, producing $2N+1$ terms to compute the indel transition probabilities [Thorne et al. 1991]. Where N random variable with rate μ_{TKF91} would determine the deletion fate and $N+1$ (one immortal link) with the exponential rate λ_{TKF91} for insertion for each character in the sequence.

The value of the smallest of these random variables determines the winner and the character of the next event in the process. Figure 2.5 represents an example of the TKF91 simulation process on an AA sequence.

As a result, the TKF91 model made the computation of the joint probability of MSA and tree possible in exponential time [Bouchard-Côté and Jordan 2013, Lunter et al. 2003]. Therefore, this model is typically implemented only in the Bayesian Markov Chain Monte Carlo (MCMC) packages. To model long indels, a closely related model, also known as TKF92, was developed assuming indivisible fragments on which the indel process acts rather than sequence residues [Thorne et al. 1992]. The main problem for the TKF92 model, similar to TKF91, is the computational cost (i.e., time and memory).

2.5.2.2 Poisson indel process model

By modifying the TKF91 model, Bouchard-Côté and Jordan [2013] introduced a single-character Poisson indel process (PIP) to describe indel and substitution of sequences using string-valued CTMC. The procedure of the PIP on a sequence of length N consists of two steps: First, the type of the next event (insertion, deletion or substitution) is determined by $2N+1$ exponential random variables (as a modified version of TKF91). Second, the exact event is determined

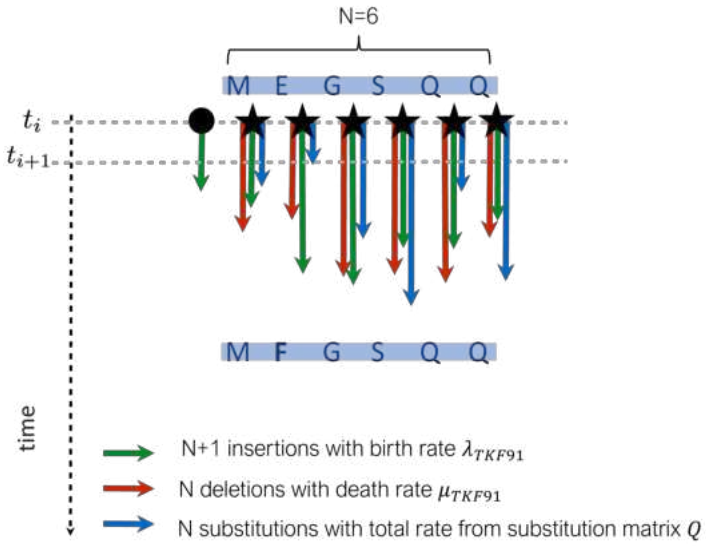


Figure 2.5: An AA sequence evolved under TKF91 model. ★ represents the normal link, and • represents the immortal link, then the AA sequence MEGSQQ could be depicted as the first line. As an example of an alignment simulation or transition path from the first sequence to the second sequence using $3N+1$ random variable of the model. Notice that the process runs on the links, not characters.

based on the type of evolutionary event and waiting time $\Delta t = t_{i+1} - t_i$ (evolution time).

The PIP is a modification of the TKF91 model, allowing the computation of the joint likelihood of tree and MSA in linear time. The stochastic process is very similar to the TKF91 except for removing the dependence of insertion rate to the sequence length [Bouchard-Côté and Jordan 2013]. Instead of having $3N+1$ exponential random variable, PIP has $2N+1$ random variables to determine the fate of the next event and the waiting time. This process is defined on the characters with N variable for substitution, 1 with rate λ_{PIP} for insertion, and N with rate μ_{PIP} for deletion. This simple modification of the insertion rate made the likelihood computation linear.

The mechanism for selecting the next mutation event (winner) is the same as TKF91, but the general result will differ. Assuming that at some point in time t_i, a sequence has length N. In the TKF91 process, the sequence stays unchanged for a random interval of time Δt, and after this interval, a single random mutation event (substitution, insertion, or deletion) alters the sequence. The value of the smallest of these random variables determines the time Δt. The winner event determines the nature of the event at time Δt [Bouchard-Côté and Jordan 2013].

The waiting time for a potential insertion event is exponentially distributed with rate $\lambda_{\text{PIP}} > 0$. The waiting times for N potential deletion events are independently and identically exponentially distributed with rate $\mu_{\text{PIP}} > 0$. The location of the deletion event is sampled from N since

there is one random variable for the deletion of each site. The waiting times for N potential substitution events are independently and exponentially distributed with rates based on the substitution rate matrix Q.

After determining the type of event, in the second step, if the event is an insertion, the insertion location is sampled uniformly from the $N+1$ possible insertion positions, and the new character is generated from a multinomial distribution with parameter π, the equilibrium distribution. If the next event is a deletion, the character associated with the minimum N deletion random variable is deleted. If the next event is a substitution, the character is randomly generated from the multinomial distribution from a row of the substitution rate matrix Q.

The driving force of each PIP event consists of two sub-processes: insertion and deletion-substitution sub-process. The former is the Poisson process defined on the phylogeny τ results in insertion point i, while the latter is substitution-deletion CTMC yields paths along sub-tree τ_i rooted at the insertion point in which a single-character either mutates or is deleted [Bouchard-Côté and Jordan 2013].

Therefore, we define the homology path H_i as a single-character site generated by substitution-deletion along the phylogeny. If the insertion point is not the root ($i \neq \Omega$), we set the value of the ancestor to the gap character. We can construct an MSA M from homology paths by extracting characters at leaves and arranging them in a column. Then, we can remove all the columns which consist only of gap characters.

From the computational viewpoint, PIP is a single-character reversible evolutionary model that computes the joint probability of MSA and the tree in linear time. PIP includes gap-state (as a result of deletion) as an additional character in the standard CTMC. This model assumes independence between sites, constant insertion rate over time, homogeneous patterns of substitutions and deletions over time, and time reversibility. Proof of time-reversibility has been provided in the original publication. This provides the convenience of an arbitrary choice of the root of a phylogenetic tree.

Most recently, other evolutionary models have been developed to model substitution and indel for the MSA tasks. However, these are also computationally challenging and of little practical use, and thus, are beyond the scope of this study, for example, see [De Maio 2021].

2.6 Ancestral sequences reconstruction

A fundamental challenge in evolutionary biology is understanding how the traits we observe today in different species evolved from some common ancestor [Gascuel and Steel 2014]. ASR

methods provided the means to solve this challenge, aiming to find the ancestral sequence that maximizes the conditional probability of the observed (extant/descendant) sequences, phylogenetic tree τ, and evolutionary model of sequences. In other words, ASR is an inference of character state at any internal node of the phylogenetic tree τ, maximizing a scoring function (usually a phylogenetic likelihood function). ASR could be useful in various applications, from protein engineering [Spence et al. 2021] to studying the progression of HIV [Ross et al. 2006] and vaccine design [Gaschen et al. 2002]. This is possible by restoring extinct proteins in the laboratory and studying their biochemical properties.

There are two main paradigms for ASR: maximum parsimony (MP) [Fitch 1971] and probabilistic-based approaches, which include empirical Bayes (EB) using maximum likelihood estimation and hierarchical (full) Bayesian reconstructions.

Historically, ASR inference was first done using the maximum parsimony (MP) method, which was implemented in the Fitch algorithm [Fitch 1971]. According to the parsimony criterion, the inference objective is based on the minimum amount of changes at each site [Brown et al. 1982]. MP reconstruction suffers from serious limitations, including producing multiple parsimonious ancestral sequences rather than just one. In MP setting, there is no way to choose between equally parsimonious solutions.

Parsimony methods do not use a model of molecular evolutionary change, so these methods generally fail to properly consider biased substitution patterns between characters (nucleotides or AAs) and different branch lengths in the tree [Yang et al. 1995]. Therefore, probabilistic model-based approaches have largely superseded parsimony-based methods [Yang et al. 1995, Pupko et al. 2000, Yang 2007]. Notably, unlike MP, probabilistic models not only provide us with the most probable ancestor (MPA) but also provide an estimate of the confidence score for each inferred ancestral character, usually expressed as a probability profile. Conversely, one advantage of most MP reconstructions is their linear time complexity concerning the number of sequences. However, new optimization and algorithmic solutions significantly reduce the time complexity of existing statistical techniques (for e.g., [Felsenstein 1981, Pupko et al. 2000]). This thesis focuses on likelihood-based ASR, a statistical approach, as current solutions are computationally time-efficient.

As in other phylogenetic applications, we use the CTMC to model substitutions, where each character state is a random variable in the ASR estimation. It is important to mention that we do not compute the likelihood of ancestral states as ML is a parameter estimation technique. Given a fixed model of substitution (and indel), both ancestral sequences and alignments are considered random variables, not model parameters. To infer the ancestral sequence, we should calculate the conditional (posterior) probability of ancestral states given the data [Yang 2014,

p. 126]. In this task, we are trying to maximize $Pr(A|D)$ where D are the data, including the states at the tips, and A are the states for internal/ancestral nodes (more details are provided in section 2.6.1).

The empirical Bayes (EB) method (in contrast to the Bayesian approach or full Bayesian) computes posterior probabilities without considering the uncertainty of parameters. The prior knowledge is simply replaced by the equilibrium (stationary or background) frequency of the CTMC process to reflect the uncertainty of the MPA. In this thesis, our new method is also referred to as the "likelihood method" of ancestral reconstruction as used in the current literature.

Before ASR, we first need to estimate the MSA, tree, and evolutionary model parameters from input sequences. Then, we infer the ancestral states using EB, with these parameters fixed. Model parameters such as tree and substitution rates are often estimated using ML. This approach cannot fully consider the uncertainty in the MPA while introducing error to the inference [Pupko et al. 2007]. The full Bayesian approach allows us to overcome this issue.

Maximum a posteriori (MAP) is an alternative to the ML with EB to reconstruct a phylogeny history. Bayesian methods calculate posterior probabilities by considering the uncertainty of the inference using the prior estimate by Markov Chain Monte Carlo (MCMC). The Gibbs sampling algorithm, a case of the Metropolis–Hastings algorithm, deploys MCMC to approximate the prior distribution of parameters [Westesson et al. 2012]. Although this method has some advantages over the two preceding approaches, in some contexts, EB is sufficient in terms of accuracy, while prior computation would not contribute much to the final inference. Despite the computational complexity of the MCMC procedure, it will allow us to deploy a complex model of substitution and indel. Bayesian paradigm is usually used for relatively small datasets. From this point, parsimony and Bayesian ASR methods are not discussed further in this thesis.

2.6.1 Likelihood method of ancestral character states

Likelihood based estimation was introduced for ASR tasks [Koshi and Goldstein 1996, Yang et al. 1995] to overcome many shortcomings of parsimony-based methods [Fitch 1971]. Unlike MP, the likelihood method takes into account branch lengths and model parameters such as substitution rates.

In the ASR method presented in this thesis, each position in the sequence is considered individually rather than estimating the whole ancestral sequence at once. We, therefore, do a single-character reconstruction, also known as the ancestral states reconstruction. This terminology uses the site-independence assumption of a CTMC, which is a standard assumption in the field of computational molecular evolution.

To formulate the problem, we aim to compute the ancestral states given the homologous sequences in an MSA M, species relationship expressed in a phylogenetic tree τ with branch lengths $\mathscr{E}$ and model parameters θ (substitution model and other parameters). Let X denote the vector of internal nodes, and $x \in \mathrm{X}$ is a single internal node. A_x^m is the ancestral character state at node x and site m. Inferring ancestral character state with the help of Bayes theorem usually relies on the computation of conditional probability $Pr(A_x^m = j|\mathrm{M}^m, \tau, \mathscr{E}, \theta)$, where j could be any arbitrary character state in the alphabet $j \in \Sigma$. Moreover, we define M^m as the mth column in the MSA matrix M while M_n^m is the mth site in the nth sequence.

$$Pr(A_x^m = j|\mathrm{M}^m, \tau, \mathscr{E}, \theta) \propto \begin{cases} Pr(\mathrm{M}^m|\tau, \mathscr{E}, \theta, A_x^m = j) \times Pr(A_x^m = j) & \text{if } x = \Omega \\ Pr(\mathrm{M}^m|\tau, , \mathscr{E}, \theta, A_x^m = j) & \text{o.w,} \end{cases} \tag{2.5}$$

$Pr(A_x^m = j)$ is approximated by the equilibrium frequency of state j. $Pr(\mathrm{M}^m|\tau, \mathscr{E}, \theta, A_x^m = j)$ is the conditional probability refers as likelihood function ($L_x(j)$) of the model given that state j is observed at node x. This conditional probability is estimated using Felsentein's recursion (pruning) algorithm instead of expensive marginalization over all the possible alphabet combinations for the ancestral node ($|\Sigma|^{l-1}$, where $l-1$ is the total number of internal nodes-X for a rooted binary tree τ).

Felsentein used the nesting rule, also known as Horner's rule, which was invented by the Chinese mathematician Zhu Shijie, to reduce the computation. Consequently, we compute ancestral states at a node only after doing so for all its descendant nodes [Yang 2014, p. 104]. Felsentein's pruning is a variant of the dynamic programming (DP) algorithm. Using the pruning algorithm, we compute the conditional probability of observing data at the tips that are descendants of node x, given that the alphabet at the descendant of node x is M^m denoted by vector $L_x(.)$ of length $|\Sigma|$. Notice that $L_l(\sigma) = 1$ for the tip nodes $l \in \mathscr{L}$, if σ is the observed alphabet or 0 otherwise. If the internal node x has two leaf children y and z denoted by set $S = \{y, z\}$, the conditional probability is:

$$Pr(\mathrm{M}^m|\tau, \theta, A_x^m = j) = \prod_{s \in S} \Big[\Sigma_{k \in \Sigma} Pr(\mathrm{M}_s^m|\tau_s, \mathscr{E}_s, A_s^m = k) \times Pr(A_s^m = k|A_x^m = j, \mathscr{E}_s) \Big], \tag{2.6}$$

Where M_s^m is the mth site character in the sth sequence corresponding to leaves of the tree rooted at x. x is the parent of $s = \{y, z\}$ and τ_s is the subtree rooted at s with branch $\mathscr{E}_s$ connecting s to its immediate parent x. $Pr(A_s^m = .|A_x^m = ., \mathscr{E}_s)$ is the transition probability under CTMC along the node x and its children also dented by $P_{xs}(\mathscr{E}_s)$ in the literature. Therefore, the probability

$L_x(j)$ of observing all descendant tips of node x is equal to the product of the probability of observing data at the descendant tips of node x, i.e., nodes y and z.

The equivalent likelihood function is:

$$L_x(j) = \left[\Sigma_y P_{xy}(\mathscr{E}_y) L_y(j)\right] \times \left[\Sigma_z P_{xz}(\mathscr{E}_z) L_z(j)\right] \tag{2.7}$$

where $P_{\cdot}(\mathscr{E})$ is the transition probability of the CTMC computed using θ (substitution model and tree). After visiting all the nodes, the likelihood of that data at the specific site is $\sum_{\sigma\in\Sigma} \pi_\sigma L_\Omega(\sigma)$ where π_σ is assumed to be the prior given by equilibrium frequency. We use four species provided in Figure 2.3 as an example. Figure 2.6 illustrates the pruning algorithm used for the ASR task.

Consequently, the inferred ancestral character $\hat{j}$ is

$$\hat{j} = \arg\max_j (Pr(A_x^m = j | \mathrm{M}^m, \tau, \mathscr{E}, \theta)), \tag{2.8}$$

The EB method usually uses Felsentein's recursion algorithm [Felsenstein 1981], a DP technique [Cormen et al. 2022, p. 359]. The pruning algorithm considers each node separately, while another layer of DP could infer the combination of all ancestral node characters that maximizes the joint posterior probability [Pupko et al. 2000].

In his book [Yang 2006], Yang distinguished two variations of EB reconstruction, joint and marginal. In joint reconstruction, we are interested in the most likely pathway from ancestor to descendant per site (without focusing on any specific nodes); however, in marginal reconstruction, we are interested in the most likely character state after one step toward descendants at a specified node of interest.

We can consider the joint scenario as the global view and the marginal scenario as the local view [Gascuel and Steel 2014]. Notice that joint and marginal reconstruction results are not necessarily the same [Pupko et al. 2000]. However, no benchmarking study compared results of joint and marginal reconstruction in a systematic way, across simulated and real-world data. The choice of marginal or joint reconstruction depends on the research question. In this thesis, we provide a solution for joint ASR.

This probabilistic representation was partly borrowed from [Pupko et al. 2007, Yang 2014, Oliva et al. 2019].

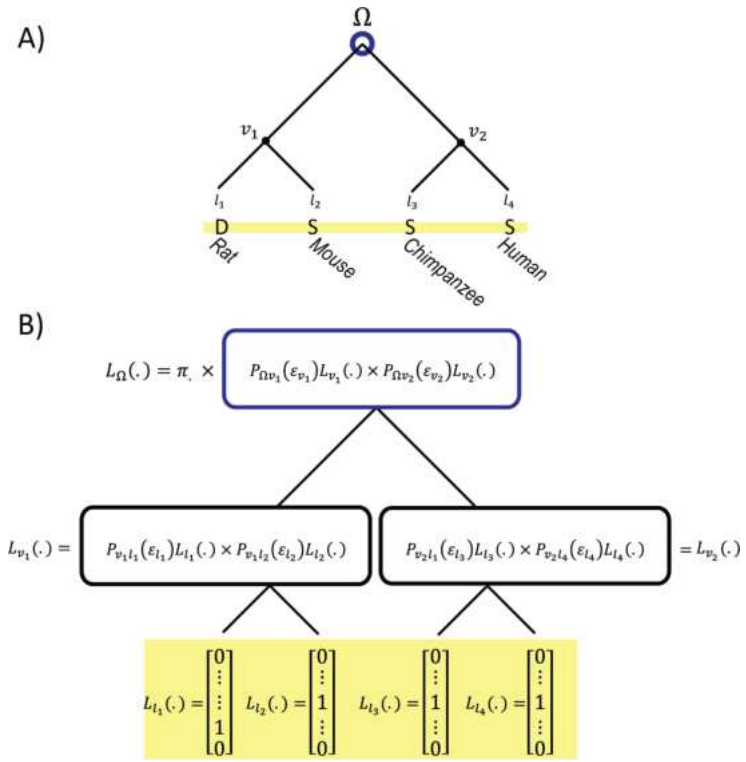


Figure 2.6: An example of Felsentein's recursion algorithm. The observed AA at the taxa is 'DSSS'. The ancestral nodes are denoted by v_1, v_2, and Ω. A) Homology path and the relationship represented in the format of a tree and MSA column. B) The detailed computation.

2.6.1.1 Joint ancestral state reconstruction using empirical Bayes

The choice of joint or marginal reconstruction depends on the study's intention (research question); however, the difference is thought to be insignificant [Pupko et al. 2000]. Although joint reconstruction is more computationally challenging than marginal reconstruction, a fast method under DP has been proposed with linear time complexity [Pupko et al. 2000]. Chapter 3 completely covers the joint ASR with substitution and indel implemented with a DP algorithm.

As Pupko et al. explained in their original FastML paper [Pupko et al. 2000], in joint reconstruction, the most likely set of ancestral states at all the internal nodes are inferred, while in marginal reconstruction, the most likely sequence at a specific internal node is inferred, averaging over all possible ancestral states at all other nodes. This means that given the phylogenetic tree and transition probabilities of moving from one state to another, we have the objective of the highest conditional probability (most likely pathway) from tips to the root, $l_1 \rightarrow v_1 \rightarrow \Omega$. In Figure 2.6, we are generally interested in the set of v_1, v_2, and Ω instead of a single node v_1.

This joint character assignment for all the internal nodes (v_1, v_2, Ω) is computed using equation 2.9 called the joint ASR. The inferred characters $\hat{p}$, $\hat{r}$, and $\hat{q}$ are

$$\hat{p}, \hat{r}, \text{ and } \hat{q} = \arg\max_{p,r, \text{ and } q}(Pr(A^m_{v_1} = p, A^m_{v_2} = r, A^m_{\Omega} = q | \mathrm{M}^m, \tau, \mathscr{E}, \theta)), \tag{2.9}$$

2.6.1.1.1 Gaps and indels in ancestral character state reconstruction The gap ("-" character) represents one or multiple insertion and/or deletion events in the homologous sequences. Most phylogenetic methods do not explicitly model indels in the evolution of a sequence, while well-known ASR methods unrealistically treat indels as "gaps", which are often represented as missing characters and removed (e.g., [Yang 2007, Pupko et al. 2000]). Current ASR methods with indels suffer from a lack of an appropriate evolutionary model of indels. Modeling indels in ASR is very important, as it allows to infer whether an ancestral state of a gap was a character or also a gap.

Kishino et al. [1990] were among the first researchers to incorporate indels into the Markov substitution model. To ML estimate the phylogeny, they proposed a simple 2-state probabilistic model for absence/presence sequence stretches equivalent to deletion/insertion events.

One approach to represent the gap-state is to consider it as an additional state in the model (i.e., the fifth state for DNA/RNA and the 21st state for AAs) [Koshi and Goldstein 1996]. This approach is obliged to consider site-independence in the phylogeny task as one needs to compute the transition probability of the gap state. This way, the insertion of two residues will be considered as two independent “character-to-gap transitions” rather than a single insertion of two characters [Pupko et al. 2007].

Edwards and Shields [2004] proposed GASP, a two-phase algorithm to infer the ASR with gaps: First, the algorithm computes the probability of gaps at each position using a two-state $0/1$ model (0 for gaps character and 1 for other characters). Once the gap/non-gap fate of the ancestral node is determined, they use an informal likelihood approach derived from empirical substitution matrices to compute the non-gaps [Pupko et al. 2007]. They show that this type of informal likelihood computation is not as accurate as ML.

Pupko and colleagues later proposed a second version of the FastML software, assuming CTMC for substitutions and a parsimony-based method called indel-coding [Simmons and Ochoterena 2000] for indels. Indel-coding unrealistic model of indels by coding them into a binary presence/absence indel matrix. The transition between these two states represents the gain and loss of genetic information. Notice that indel-coding reconstructions are not computed based on CTMC, as it is only used for substitution reconstruction. FastML applies an ML-based algorithm for binary data to compute the probability of gap/non-gap in ancestral sequences. In the non-gap position, the second algorithm infers the MPA sequence. Overall, the approaches coding gaps in a binary way (presence/absence of a character) allow deleted characters to give rise

to a new non-character, which is biologically non-realistic.

Some studies [Cartwright 2009] distinguish gaps from indels. An indel is a specific evolutionary event that adds or removes residues from a sequence, whereas a gap results from one or more overlapping or adjoining indels, making residues in one sequence nonhomologous to any residues in the other sequence. Indel mutations, unlike substitution mutations, may involve several sites that vary in length and overlap. For those ASR methods with indels, they do not distinguish between insertion and deletion events.

Conventional ASR methods struggle with incorporating indel information. The ability to accurately reconstruct long indel (i.e., multiple-character indel) is the main challenge in the existing ASR methods with indels. With the exception of hand-coded indel characters, mainstream methods for phylogenetic tree reconstruction have been refractory to the incorporation of the indel information present in the sequence data. Another question is whether a single-residue indel model could be sufficient for ASR of sequences with long indel or whether indels should be modeled with multi-residue indel events.

Chapter 3

ARPIP: Ancestral Sequence Reconstruction with Insertions and Deletions under the Poisson Indel Process

This chapter, aiming to infer ancestral sequences with indels, was published in the Journal of Systematic Biology. This method was developed based on the PIP model, which performs both biologically and computationally adequately. PIP is a single-character CTMC stochastic process defined on the phylogeny, which computes the joint likelihood of the tree and alignment in linear time.This work introduces and develops an algorithm to reconstruct the ancestral sequence of desired homologous species under the PIP model. This chapter will also cover the background to understand the PIP model and the algorithm in appendices. ARPIP is the key contribution of this dissertation, which heavily relies on the role of indels in the evolutionary history of species. The ARPIP software was implemented using C++ programming language under GNU GPLv3 license. The code and experimental data used in the manuscript, along with short documentation, are freely available on the GitHub page `https://github.com/acg-team/bpp-ARPIP`. This work was published in July 2022 at *Systematic Biology* and later printed in Volume 72, Issue 2, March 2023, Pages 307–318, `https://doi.org/10.1093/sysbio/syac05`, where I am the first and corresponding author. Following is the published version of the article.

ARPIP: Ancestral Sequence Reconstruction with Insertions and Deletions under the Poisson Indel Process

Gholamhossein Jowkar[1,2,3,*], Jūlija Pečerska [1,2], Massimo Maiolo [1,2,4], Manuel Gil [1,2], Maria Anisimova [1,2]

[1] *Zurich University of Applied Sciences, LSFM, ICLS, CH-8820, Wädenswil, Switzerland*
[2] *Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*
[3] *University of Neuchâtel, Institute of biology, CH-2000 Neuchâtel, Switzerland*
[4] *University of Bern, Institute of Pathology, CH-3008 Bern, Switzerland*

**Gholamhossein Jowkar, ZHAW, School of Life Sciences and Facility Management, Applied Computational Genomics Group, Schloss 1, 8820 Wädenswil, Switzerland, E-mail: jowk@zhaw.ch*

Abstract

Modern phylogenetic methods allow inference of ancestral molecular sequences given an alignment and phylogeny relating present day sequences. This provides insight into the evolutionary history of molecules, helping to understand gene function and to study biological processes such as adaptation and convergent evolution across a variety of applications. Here we propose a dynamic programming algorithm for fast joint likelihood-based reconstruction of ancestral sequences under the Poisson Indel Process (PIP). Unlike previous approaches, our method, named ARPIP, enables the reconstruction with insertions and deletions based on an explicit indel model. Consequently, inferred indel events have an explicit biological interpretation. Likelihood computation is achieved in linear time with respect to the number of sequences. Our method consists of two steps, namely finding the most probable indel points and reconstructing ancestral sequences. First, we find the most likely indel points and prune the phylogeny to reflect the insertion and deletion events per site. Second, we infer the ancestral states on the pruned subtree in a manner similar to FastML. We applied ARPIP on simulated datasets and on real data from the *Betacoronavirus* genus. ARPIP reconstructs both the indel events and substitutions with a high degree of accuracy. Our method fares well when compared to established state-of-the-art methods such as FastML and PAML. Moreover, the method can be extended to explore both optimal and suboptimal reconstructions, include rate heterogeneity through time and more. We believe it will expand the range of novel applications of ancestral sequence reconstruction.

Keywords: joint ancestral sequence reconstruction, ancestral sequences, maximum likelihood, indel, phylogeny, dynamic programming, Poisson indel process, SARS-CoV, evolutionary stochastic process.

Phylogenetics is a wide research field with a variety of applications ranging from reconstructing the tree of life to investigating ongoing epidemics. Phylogenetic trees provide insight into unobservable evolutionary events in the past such as adaptation or mass extinction events. Phylogenetic inference can be divided into several interrelated tasks including sequence alignment, phylogeny estimation, detection of selection, and ancestral sequence reconstruction (ASR). ASR aims to infer the likely ancestral sequences for a set of existing homologous sequences.

ASR allows researchers to pursue a wide range of topics from determining the origins of life or epidemics to developing personalised medicine [Pagel 1999, Liberles 2007]. For example, the functionality of ancient genes can be investigated by reconstructing and synthesising the genetic material inferred by ASR [Thornton 2004]. Such analyses can help us understand the mechanisms underlying adaptation and speciation processes, inspiring new approaches for protein engineering [Chang et al. 2005] and drug design [Zakas et al. 2017]. ASR can be used to study epidemiological origins of pathogens, particularly in light of recent coronavirus pandemics [Pagel 1999, Brintnell et al. 2021, Starr et al. 2022].

State-of-the-art likelihood-based ASR methods use Markov processes to model character substitutions through time. Such models account for various biases in character substitution, as well as divergence represented by evolutionary time [Yang et al. 1995, Yang 2007, Pupko et al. 2000]. However, Markov models of molecular evolution do not include insertions or deletions (indels) as part of the evolutionary process, meaning that methods relying on these models have to treat gap characters separately. Most of the ASR methods adopt one of two pre-processing approaches. They either treat gaps as missing/ambiguous data or remove gap characters entirely. However, indels represented by gaps carry an important evolutionary signal [Dessimoz and Gil 2010], and are in fact a major driving force of genomic divergence [Tao et al. 2007]. Therefore, methods that model indels explicitly have a clear advantage over methods that do not. Up to this point, most existing frequentist algorithms do not include indel modelling except for two methods, Ancestors [Diallo et al. 2009] and FastML [Ashkenazy et al. 2012]. Ancestors has exponential computational complexity and therefore has not been widely adopted by users. FastML handles indels using a heuristic approach called indel-coding. The method relies on the linear time complexity algorithm [Pupko et al. 2000] for joint maximum likelihood (ML) ASR using dynamic programming (DP). FastML makes the analyses of large datasets tractable. Currently, it is provided as a web service [Ashkenazy et al. 2012]. While the results of indel-coding can be interpreted from an evolutionary standpoint retrospectively, the approach does not, however, include an explicit evolutionary indel model. All things considered, most methods rely on standard models of sequence evolution without indels which is an issue that can only be resolved by including character and indel evolution in a single model.

Two pioneering mathematical models describing the evolution of indels are TKF91 and TKF92

[Thorne et al. 1991; 1992]. However, the computation of marginal likelihood under these models has exponential time complexity, rendering the methods relying on these models extremely computationally intensive, and making inference under these models unrealistic on large datasets. More recently, [Bouchard-Côté and Jordan 2013] proposed the Poisson Indel Process (PIP) model which is based on TKF91. PIP describes insertions by a Poisson process defined on the tree topology, while substitutions and deletions are described by a continuous-time Markov process where deletions are modelled as an absorbing state. The assumption of independence between insertion and substitution/deletion enabled a major computational improvement over the previous models [Bouchard-Côté and Jordan 2013]. In PIP, the insertion rate is also independent of the length of a sequence, which is a realistic assumption based on the data that is most commonly analysed [Bouchard-Côté 2010, p. 93]. In contrast to TKF91, the PIP model allows to compute marginal likelihoods in linear time with respect to the number of sequences, which enables a variety of phylogenetic applications (e.g., Maiolo et al. [2018]).

In this study, we use the PIP model for joint reconstruction of ancestral character states including insertions and deletions. Our method ARPIP (Anscestral Reconstruction under PIP) is implemented in the ML framework, i.e., we use an empirical Bayesian approach with ML estimates. Given a multiple sequence alignment (MSA) and a phylogenetic tree, we first use PIP to infer insertion and deletion points on the tree. Insertion and deletion points are the specific locations on the phylogeny where the events have happened. Next, we extract a subtree rooted at the insertion point and pruned by the deletion points. Finally, we reconstruct ancestral states on the extracted subtree using a modified version of Felsenstein's recursion [Felsenstein 1981], similar to the FastML algorithm [Pupko et al. 2000]. In the following we describe the method in detail, validate it by simulations, and demonstrate its performance in simulations and on a real dataset.

3.1 Materials and methods

ARPIP consists of two main algorithms: indel point inference and ancestral character inference.

The IndelPoints algorithm infers the most likely indel points for each site m of the given alignment (Appendix 3.1.2). It traverses the tree in post-order and evaluates a set of possible indel scenarios for each node in the tree. A particular indel scenario defines a homology path $\mathscr{H}$. A homology path contains a single insertion point and a number of deletion points consistent with the input MSA. IndelPoints finds the most likely indel scenario by maximizing the probability of $\mathscr{H}$ given m. The maximisation is simplified by reducing the MSA to gap and non-gap states, and ignoring the substitution history without changing the result of the computation. This al-

lows us to avoid matrix exponentiation, which is computationally expensive but necessary for the full likelihood computation.

Similar to the recursive likelihood computation, we traverse the tree and evaluate all the possible indel scenarios, selecting the best one at each node. At the tree root, we select the best homology path over the whole tree based on the best paths selected in the child nodes. For each site m, we use the inferred homology path to extract a subtree τ_m rooted at the insertion point $\mathscr{I}$ and pruned by deletion points $\mathscr{D}$, which represents the most likely indel history for the given site.

Next, we reconstruct ancestral characters on the pruned subtrees in a manner similar to FastML [Pupko et al. 2000]. For each site m, we use DP to reconstruct ancestral characters in two phases. The first phase can be seen as a modification of Felsenstein's peeling recursion for computing marginal likelihoods [Felsenstein 1981]. As in the peeling recursion algorithm, we traverse the tree τ_m in post-order, starting from the leaves upward to the root and propagate partial likelihoods. However, instead of marginalising over internal character states, for each MSA column m we store the likelihood values Lk_v and the corresponding best ancestral states CS_v for each node v. In the second phase, the algorithm traverses the tree in pre-order and for each node selects the ancestral character A_v with the highest conditional probability.

3.1.1 Preliminaries: the PIP model

The PIP model describes the evolutionary process of substitutions, insertions and deletions along the branches of a phylogenetic tree τ. Here we include the basic description of the process, additional information on the PIP likelihood is available in Appendix 3.7.1 and a detailed description of PIP can be found in [Bouchard-Côté and Jordan 2013].

Let $\tau = (\mathscr{V}, \mathscr{E}, \mathrm{b})$ represent a rooted binary phylogenetic tree, where set $\mathscr{V}$ is the set of all vertices of the tree, $\mathscr{E}$ is the set of all tree branches ($\mathscr{V} \times \mathscr{V}$) and b refers to the branch lengths in units of time (measured in expected substitutions and deletions per site).

The observed sequences are strings of characters from an alphabet Σ, which can be nucleotides, amino acids or codons. The N observed sequences at the leaves of τ are denoted by set $\mathscr{L} \subset \mathscr{V}$, whereas set $\mathscr{V} \setminus \mathscr{L}$ is the set of $N-1$ internal vertices. The root, the most recent common ancestor of all leaves, is labelled by Ω. The branch length $\mathrm{b}(v)$ associated with node ($v \in \mathscr{V}$) spans from v to its parent vertex $\mathrm{pa}(v)$ (see Fig. 3.1).

PIP is parameterised by insertion rate λ and deletion rate μ, with the process running over tree topology τ. For every node $v \in \mathscr{V}$, the probability of inserting a single character on edge $e = (\mathrm{pa}(v) \to v)$ is proportional to the branch length and defined by $\iota(v)$ (see Appendix 3.7.1). Similarly, the survival probability for a character inserted on edge e is $\beta(v)$ (see Appendix

3.7.1). Additionally, we define the pure survival probability $\zeta = \exp(-\mu \mathrm{b}(v))$ associated with node v as if the character was already present at the parent node $\mathrm{pa}(v)$ [Maiolo 2019]. Point substitutions and deletions are modelled by a continuous-time Markov process on $\Sigma_\varepsilon = \Sigma \cup \{\varepsilon\}$, where ε denotes the gap symbol. The generator matrix Q could be any arbitrary reversible substitution model, e.g. WAG for amino acids [Whelan and Goldman 2001], or K80 for nucleotide data [Kimura 1980]. Accordingly, the extended generator matrix is denoted by Q_ε and the extended quasi-stationary distribution is $\pi_\varepsilon = [\pi, 0]$ [Bouchard-Côté and Jordan 2013].

Let $\mathscr{G}$ define the site-specific set of all potential deletion points on the tree. $\mathscr{G}$ consists of all leaves with a gap at the respective site, and of all the internal nodes whose all descendant leaves have a gap at that site. Next, consider the subset $\mathscr{S}$ of leaves that have a non-gap character, $\mathscr{S} = \{v \in \mathscr{L} : m_v \neq \varepsilon\}$. Given the set $\mathscr{S}$, we define the set $\mathscr{A}$ of potential insertion points to include all nodes that are ancestral to all the leaves in $\mathscr{S}$ (see Fig. 3.1). In general, we compute the probability $p(m)$ of each individual MSA column by marginalizing over all possible homology paths underlying that MSA column (see Appendix 3.7.1) based on their homology path probabilities f_v.

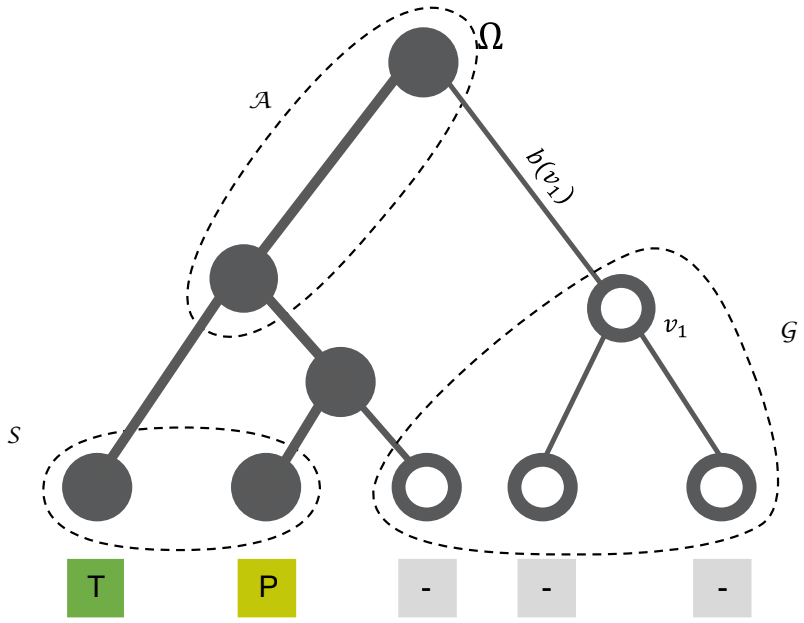


Figure 3.1: The phylogenetic tree τ rooted at Ω. $b(v_1)$ represents the branch length from Ω to v_1. The leaves of the tree show a single column of the MSA including gaps as an additional character state. The set $\mathscr{S}$ is defined as all leaves with a character in the given column (not a gap). The set of potential insertion nodes $\mathscr{A}$ contains the nodes ancestral to all nodes in $\mathscr{S}$. Finally, the set of potential deletion nodes $\mathscr{G}$ is defined as all nodes which are either a leaf with a gap in the given column, or a node whose both children are in $\mathscr{G}$.

3.1.2 Inferring the indel points

We propose a progressive algorithm to infer the most likely indel points (homology path) on the tree under the PIP model. For each site, we progressively find the best partial homology path (constrained by a subtree) and build on the intermediate results to get the most likely indel history on the whole tree. Since we search only for the most likely homology path, we compute a simplified likelihood function which accounts only for insertions and deletions and ignores substitutions.

Under PIP, two mutually exclusive node sets exist on the tree topology τ: the set of nodes where the character has gone extinct and the set of nodes where the character has definitely survived. The first is the set of potential deletion nodes $\mathscr{G}$, defined in the previous section. The second contains all the remaining nodes in the tree $v \in \mathscr{V} \setminus \mathscr{G}$ ($v \notin \mathscr{G}$). A node $v \notin \mathscr{G}$ may also be a potential insertion location, i.e. $v \in \mathscr{A}$ (see Fig. 3.2 and Appendix 3.7.2 for the detailed description). While the set $\mathscr{A}$ may contain multiple nodes, a homology path can only have a single insertion location. Consequently, each node in $\mathscr{A}$ is associated with a single homology path with the highest probability. This implies that when computing the probability of a homology path for node $v \in \mathscr{A}$, it is treated as the only potential insertion location, while all other nodes are treated as regular nodes in the tree. Notably, one cannot simply select the node with the highest insertion probability $\iota(v)$, as the probability of any given homology path also depends on the survival/extinction of the site in the children. Even though we separately describe the treatment of the two node types, all the necessary computation can be done in a single post-order traversal of the tree.

For each node v in the tree, we first compute f_v, the conditional probability of the deletion/substitution process, assuming that the character exists in v. We compute f_v for the most likely deletion scenario in this subtree rather than marginalise over all possible deletion locations. We also compute p_v, the conditional probability of the homology path assuming that the character was inserted at node v. A character necessarily has to be inserted at one of the nodes $v \in \mathscr{A}$, which means that the probability will be non-zero only for the potential insertion nodes.

In the progressive algorithm we maintain several node sets that are needed to define the most likely homology path per node. Let $\mathscr{I}_v$ denote the set of insertion points for the subtree rooted at v. Then $\mathscr{I}_v = \emptyset$ for $v \notin \mathscr{A}$ and $\mathscr{I}_v = \{v\}$ for $v \in \mathscr{A}$. Similarly, $\mathscr{D}_v$ denotes the set of deletion points for the subtree rooted at v.

For each node v we store the locally optimal homology path $\mathscr{H}_v = \{\mathscr{I}_v, \mathscr{D}_v\}$ for the subtree rooted at v. Once we reach the root node Ω, all possible insertion locations would be considered, and the one with the highest probability is selected among those. At this point, the best

homology path $\mathscr{H}_{\arg\max(p_v)}$ (defined by the highest conditional probability p_v) is used to extract the subtree τ_m rooted at $\mathscr{I}_{\arg\max(p_v)}$ and pruned by $\mathscr{D}_{\arg\max(p_v)}$, which represents the best possible indel points for the MSA column m. We will use τ_m to infer ancestral character states for column m. The IndelPoints algorithm is presented in Figure 3.2 and the pseudocode for the algorithm can be found in the Appendix 3.7.2.

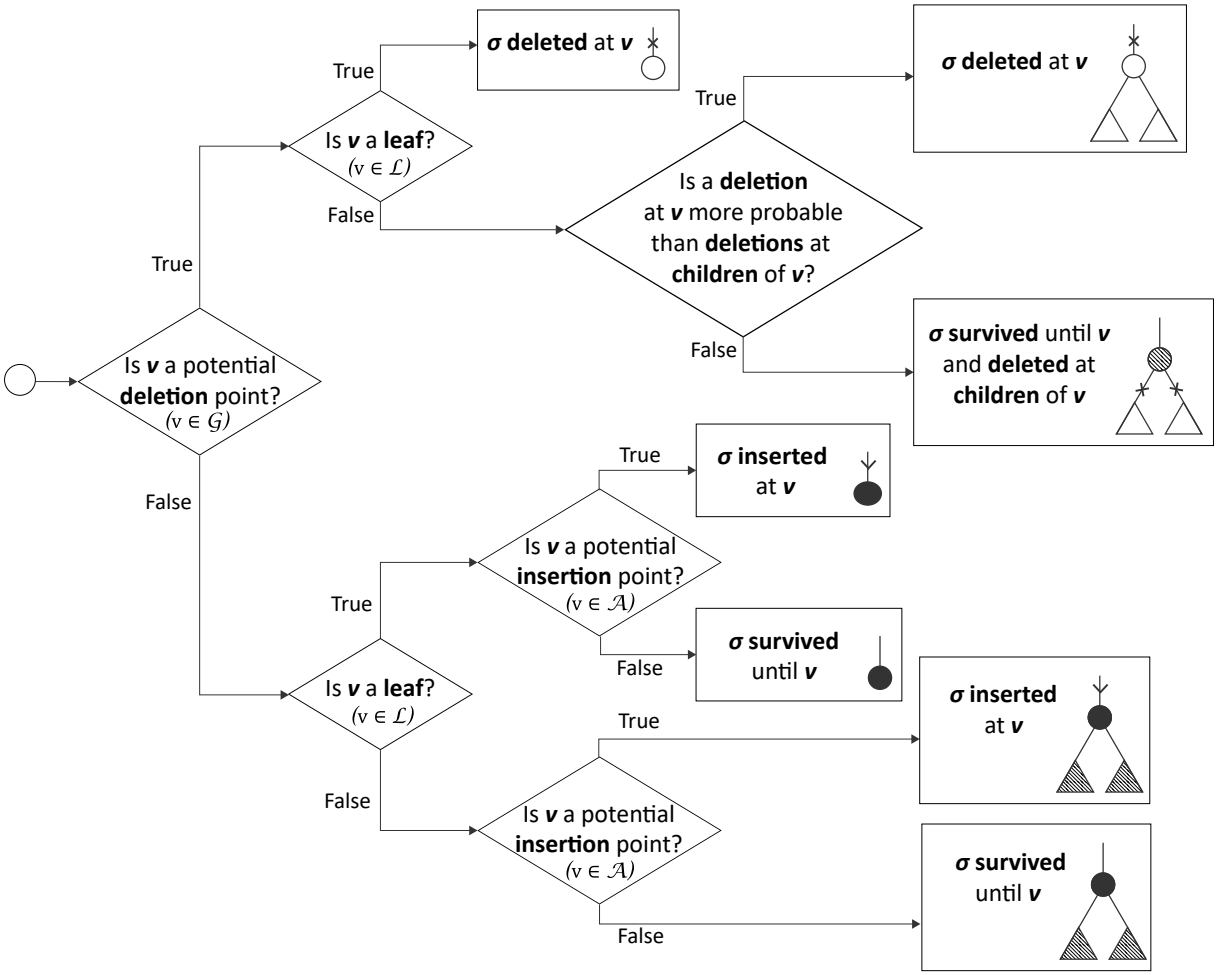


Figure 3.2: Overview of the IndelPoints algorithm. The tree is traversed in post-order to infer the most likely homology path progressively using the predefined sets: $\mathscr{L}$ the set of all leaves, $\mathscr{A}$ the set of potential insertion points, and $\mathscr{G}$ the set of potential deletion points. Here, σ represents the character in focus and v is the node visited during the tree traversal.

3.1.3 Dynamic programming joint ancestral sequence reconstruction

Our method performs ASR in a manner very similar to FastML [Pupko et al. 2000] with two crucial differences. First, we only work on a subtree τ_m of the original tree τ, which limits the reconstruction to the most probable insertion location at this site. This means we do not reconstruct any ancestral states where there were none. Second, to appropriately account for character deletion, the ancestral reconstruction is done using the PIP substitution rate matrix

Q_ε.

The joint ASR method under PIP given column m and the pruned rooted phylogenetic subtree τ_m consists of two steps. The first step is to compute the partial likelihood values on subtree τ_m with the modified version of Felsenstein recursion algorithm, where both likelihood values and their corresponding ancestral character states are stored. The second step is to reconstruct the character states by picking the character with highest conditional probability. The recursive algorithm for joint ASR is shown in Appendix 3.7.3 together with the newly defined pseudocode for the procedure.

3.2 Results

Three datasets were used to evaluate and illustrate our method. The first dataset was simulated under the PIP. This dataset allows to evaluate the performance of ARPIP under the true model. Given the true simulated trees and MSAs, both the homology path inference and ASR were evaluated.

The second dataset was used to evaluate the performance of ARPIP for sequences with long indels. The data was generated by INDELible [Fletcher and Yang 2009] with two different settings using the same trees as for the PIP simulations. For this dataset, the ancestral sequences are also known and can be used for evaluation. However, the PIP parameters for this dataset must be inferred. As INDELible does not provide a comprehensive description of indel events on the phylogeny, we used these simulations to evaluate ancestral state inference.

The third dataset is a small coronavirus sample which was extracted from Uniprot [Bateman et al. 2020]. With this dataset, we aimed to provide a showcase of the method.

When analysing both INDELible and real-life data, we first have to infer the PIP parameters, i.e. λ and μ, given an MSA and a tree. This computation was done based on Brent's optimization method [Brent 1973], optimizing one parameter at a time until convergence. For all examples, the protein substitution model used is WAG [Whelan and Goldman 2001].

Note that ARPIP was developed for rooted trees. If an unrooted tree is provided, ARPIP uses mid-point rooting method to root the tree. Furthermore, ARPIP can perform ASR without a provided tree. The user can select from established fast methods like neighbor joining, BioNJ, UPGMA, and WPGMA to estimate the tree from the input MSA.

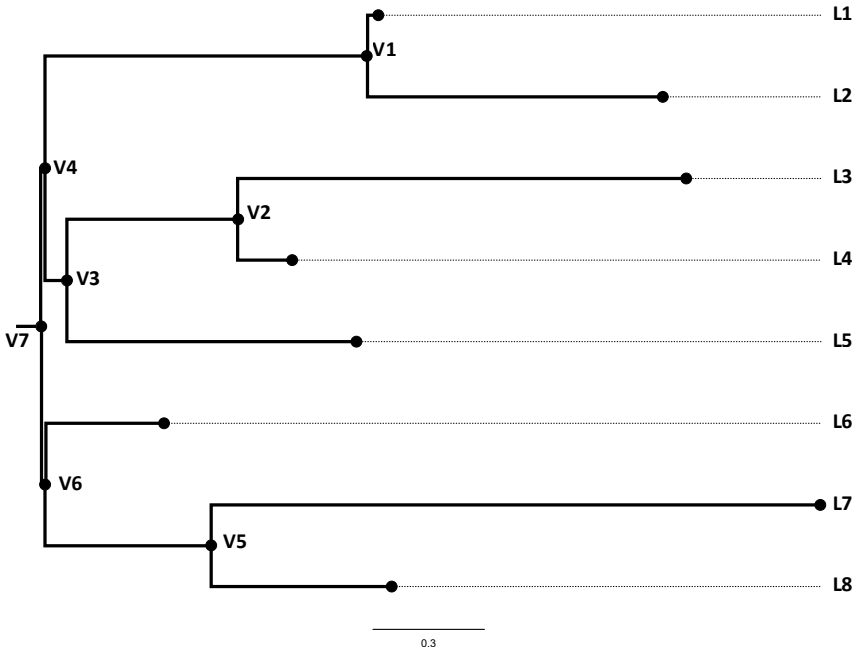


Figure 3.3: An example tree from the dataset generated by the PIP simulator.

3.2.1 Data simulated under PIP

The simulated sequences are given as input to ARPIP along with the true model parameters so that we only have to estimate the ancestral state values. The simulated dataset contains 100 MSA/tree replicates with their corresponding evolutionary events. Each replicate was simulated using an 8 taxa tree with a topology sampled from the uniform distribution and branch lengths sampled from an exponential distribution with the rate $\rho = 2$, where ρ is a proxy for phylogenetic divergence. One of the simulated trees is shown in Figure 3.3. On average, the branch lengths of the simulated trees were 0.45 units of time, ranging from minimal branch length of 0 and maximum branch length of 3.23. For the simulations, we set the deletion rate $\mu = 0.1$ and the insertion rate $\lambda = 10$ for PIP.

3.2.1.1 Analysis of the PIP simulated dataset

In order to assess the accuracy of ARPIP, we independently evaluated each inference step, IndelPoints and the joint ASR (see Tab. 3.1). To asses the accuracy of the IndelPoints algorithm, we also evaluated the inference of insertion and deletion events independently. As this dataset was simulated under the same model we use for inference, we used the true parameter values in the analysis without inferring them ($\mu = 0.1$ and $\lambda = 10$). This way we can evaluate the method without the additional variation of parameter inference, which is done for the other two datasets.

Among the 100 input sets, 96.69% insertion points and 95.54% deletion points were inferred

correctly. In the next step, we computed the accuracy of ASR per site. This number has been averaged over all existing sites over all MSA replicates. To evaluate the reconstructed ancestral sequences, we used three different metrics. Firstly, we counted the number of full ancestral columns that were inferred correctly, which is 60.08% for this dataset. Secondly, we counted the number of characters that were inferred correctly, which amounts to 88.14% of characters. Thirdly, we counted the number of gap characters themselves that were inferred correctly, which amounts to 99.86%.

Metric	**Accuracy (%)**
Correctly inferred insertion points	96.08 ± 2.84
Correctly inferred deletion points	95.54 ± 2.80
Correctly inferred MSA columns	60.08 ± 9.60
Correctly inferred characters including gaps	88.14 ± 3.91
Correctly inferred gap character	99.86 ± 0.26

Table 3.1: ARPIP accuracy for inference on PIP simulated data.

3.2.2 Data generated by INDELible

The data simulated by INDELible contains two sets of 100 MSA/tree replicas. Each replica was simulated using an 8 taxa tree from PIP simulations. We used the Zipfian (power law) distribution for the indel model with $a = 1.7$, to generate the samples where $a > 1$ is the exponent characterizing the distribution. Empirical estimates of value a range from 1.5 to 2 [Fletcher and Yang 2009], which prompted us to select $a = 1.7$. The maximum indel length was set to 5 to avoid MSAs with excessively long gaps. Two different indel rates of 0.01 and 0.05 were used for the simulation with INDELible.

3.2.2.1 Analysis of the INDELible simulated dataset

For this dataset, ARPIP inferred the PIP parameters λ and μ as well as ancestral character states. For the two datasets of 100 MSA/tree replicates with indel rates of 0.01 and 0.05, ARPIP correctly inferred 46.58% and 59.49% of the ancestral sites respectively. Further, ARPIP correctly inferred 83.49% and 87.93% of characters including gaps. Finally, over 99.95% of gap characters were inferred correctly for the two datasets (see Tab. 3.2).

3.2.3 Coronavirus data

The ongoing *SARS-CoV-2* pandemic strongly affects our lives, causing an immense interest for phylogenetic analyses of the relevant viral molecular sequences. Like in other coronaviruses, the spike protein in *SARS-CoV-2* is important for viral entry into host cells. It is also one of

Metric	Accuracy (%)	
	indel rate 0.01	**indel rate 0.05**
Correctly inferred MSA columns	46.58 ± 13.22	59.49 ± 10.63
Correctly inferred characters including gaps	83.49 ± 6.04	87.93 ± 4.33
Correctly inferred gap character	99.98 ± 0.16	99.95 ± 0.14

Table 3.2: ARPIP accuracy for inference on INDELible simulated data.

the major determining factors of host range [Belouzard et al. 2012, Zhou and Zhao 2020]. We therefore used this protein as an example demonstrating ancestral sequence inference.

SARS-CoV-2 is a member of the *Betacoronavirus* genus which also contains the two other recent human coronavirus strains, namely *SARS-CoV* and *MERS-CoV* [Lefkowitz et al. 2018]. For our analyses, we selected a small set of available protein sequences from this genus (see Tab. 3.3 for the exact sequence list).

Subgenus	Species	Uniprot acession number
Embecovirus	*Betacoronavirus 1*	A0A191URB2
	China Rattus coronavirus HKU24	A0A0A7UZR7
	Human coronavirus HKU1	U3NAI2
	Murine coronavirus	P11224
	Myodes coronavirus 2JL14	A0A2H4MXV6
Hibecovirus	*Bat Hp-betacoronavirus Zhejiang2013*	A0A088DJY6
Merbecovirus	*Hedgehog coronavirus 1*	A0A4D6G1A4
	Middle East respiratory syndrome-related coronavirus	K9N5Q8
	Pipistrellus bat coronavirus HKU5	A3EXD0
	Tylonycteris bat coronavirus HKU4	A3EX94
Nobecovirus	*Rousettus bat coronavirus GCCDC1*	A0A1B3Q5W5
	Rousettus bat coronavirus HKU9	A3EXG6
Sarbecovirus	*Severe acute respiratory syndrome-related coronavirus*	A0A3Q8AKM0
	Severe acute respiratory syndrome-related coronavirus 2	P0DTC2

Table 3.3: *Betacoronavirus* sequences used in the analysis.

3.2.3.1 Analysis of the coronavirus dataset

The MSAs of the coronavirus sequences were inferred using ProPIP [Maiolo et al. 2018] and PRANK phylogeny-aware webserver [Löytynoja 2014]. The total length of the reconstructed

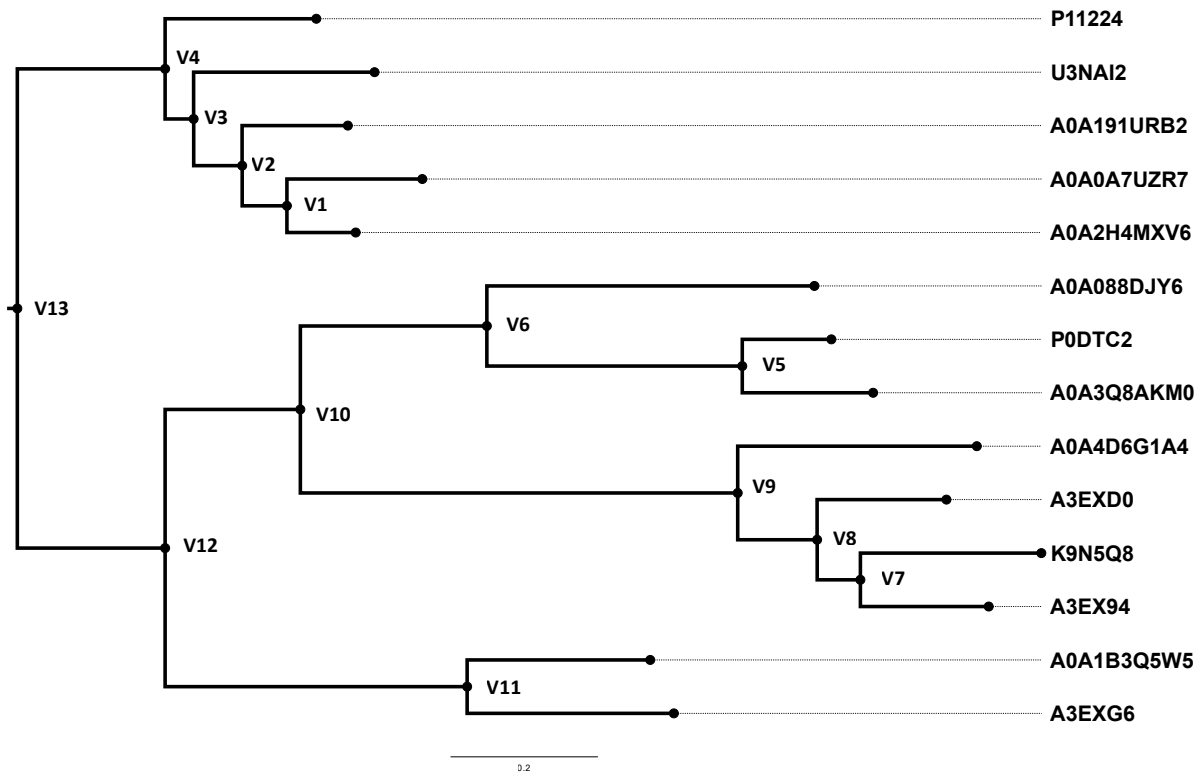


Figure 3.4: Illustration of the rooted *Betacoronavirus* phylogenetic tree which was reconstructed by PhyML 3.0 from the ProPIP alignment. Note that the original tree was unrooted which ARPIP used mid-point rooting method to make the tree rooted.

MSAs was 2002 and 1929 AAs respectively. The phylogenetic trees were reconstructed by ML in PhyML 3.0 [Guindon et al. 2010], using smart model selection on amino acids and SPR tree moves [Lefort et al. 2017] (see Fig. 3.4). Then, given an MSA and tree we inferred ancestral sequences with ARPIP. The estimated deletion rates for ProPIP and PRANK's MSAs are respectively $\hat{\mu} = 0.242$ and $\hat{\mu} = 0.210$. Figure 3.5 summarises the resulting ASR by ARPIP comparing to FastML on MSA produced by ProPIP while the results for PRANK can be found in the supplement.

3.2.4 Comparison against the state-of-the-art methods

At this moment, the two most frequently used ML joint ASR approaches are PAML [Yang 1997] and FastML [Pupko et al. 2000, Ashkenazy et al. 2012], both of which work in linear time with respect to the number of sequences. An important distinction among the methods lies in the way they handle gaps in the alignment. PAML can either ignore gaps in the alignment by removing all columns containing at least one gap character (option used here), or treat all gap characters as ambiguous. For this study, we used PAML on an MSA without any gap characters to compare the accuracy of ancestral character reconstruction. FastML webservice [Ashkenazy

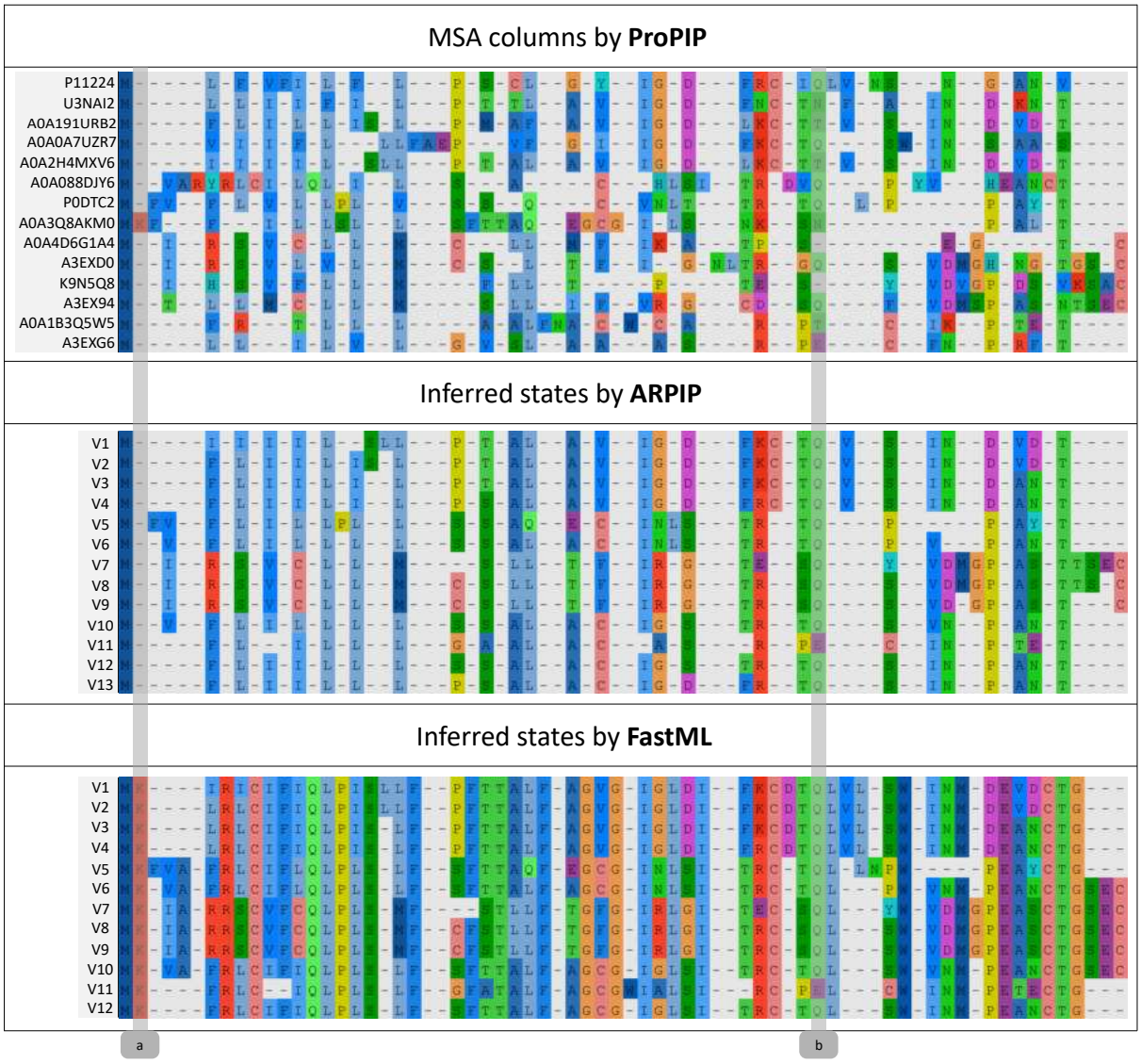


Figure 3.5: Illustration of a snippet from the *CoV* dataset containing the MSA inferred by ProPIP and the ancestral sequences predicted by ARPIP and FastML. a) The region in which ARPIP infers a very different ancestral history, probably due to inferring the insertion point prior to ancestral character inference. FastML inferred no gap in this column perhaps due to the adjacent (first) column. b) The region in which both algorithms had similar inferences of the ancestral states. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

et al. 2012] uses an ad-hoc indel-coding [Simmons and Ochoterena 2000] approach to account for indels spanning multiple adjacent characters. Indel coding is done as a separate step in the inference process, done independently from the ancestral state reconstruction. We compare the performance of ARPIP and FastML on an MSA with gaps.

On both simulated datasets, the accuracy of ARPIP appears similar to FastML (see Tabs. 3.1- 3.2 and Figs. 3.6, 3.7 and 3.8). For example, both algorithms inferred the ancestral state accurately in certain regions (e.g., Fig. 3.6a, Fig. 3.7a and d, and Fig. 3.8a) and falsely in other regions (e.g., Fig. 3.6b, Fig. 3.7b and c, and Fig. 3.8b). In region c of Figure 3.6 and region d of Figure 3.8, FastML inferred a character state even though there is no ancestral character, since the insertion happened at the leaf. In certain regions FastML could not determine which internal node had the information (e.g., Fig. 3.6d), while ARPIP was capable of determining the character position accurately. ARPIP outperformed FastML in determining the gap position in certain regions (e.g., Fig. 3.6c and d, Fig. 3.7b, and Fig. 3.8c and d). On *CoV* data, the situation was similar meaning FastML could not detect the insertion location (see Fig. 3.5a) while in conserved regions the inferred states were almost identical (see Fig. 3.5b).

We also considered scenarios without indels, allowing the evolutionary process to work only through substitutions. In this case the alignments have no gaps, i.e. no deletion ($\mu = 0$) and all insertions happen at the root of the tree. Under these conditions, all the algorithms perform reasonably as presented in Figure 3.9.

3.3 Discussion and conclusion

In this article, we present a one-of-a-kind approach for fast likelihood-based ancestral sequence reconstruction with insertions and deletions. Unlike previous approaches, our method relies on an explicit model of indel and character evolution and allows us to infer the full history of sequence evolution, including insertion and deletion points on a phylogeny. The method is implemented in the probabilistic framework and is based on likelihood calculations under the PIP model. Likelihood computations under this model have linear time complexity with respect to the number of sequences, meaning that our method is highly efficient on large datasets.

We show that on PIP simulated datasets, ARPIP correctly infers at least 95% of indel events and at least 88% of ancestral characters. On the INDELible simulated data, ARPIP correctly infers 83% and 87% of ancestral characters including gaps for low and high indel rates respectively. ARPIP also correctly places gaps in over 99.95% cases, showing the credibility of our Indel-Points algorithm. For all datasets, we illustrate the performance of our approach in comparison to FastML on alignments with gaps. In addition, we use gapless alignments to compare ARPIP

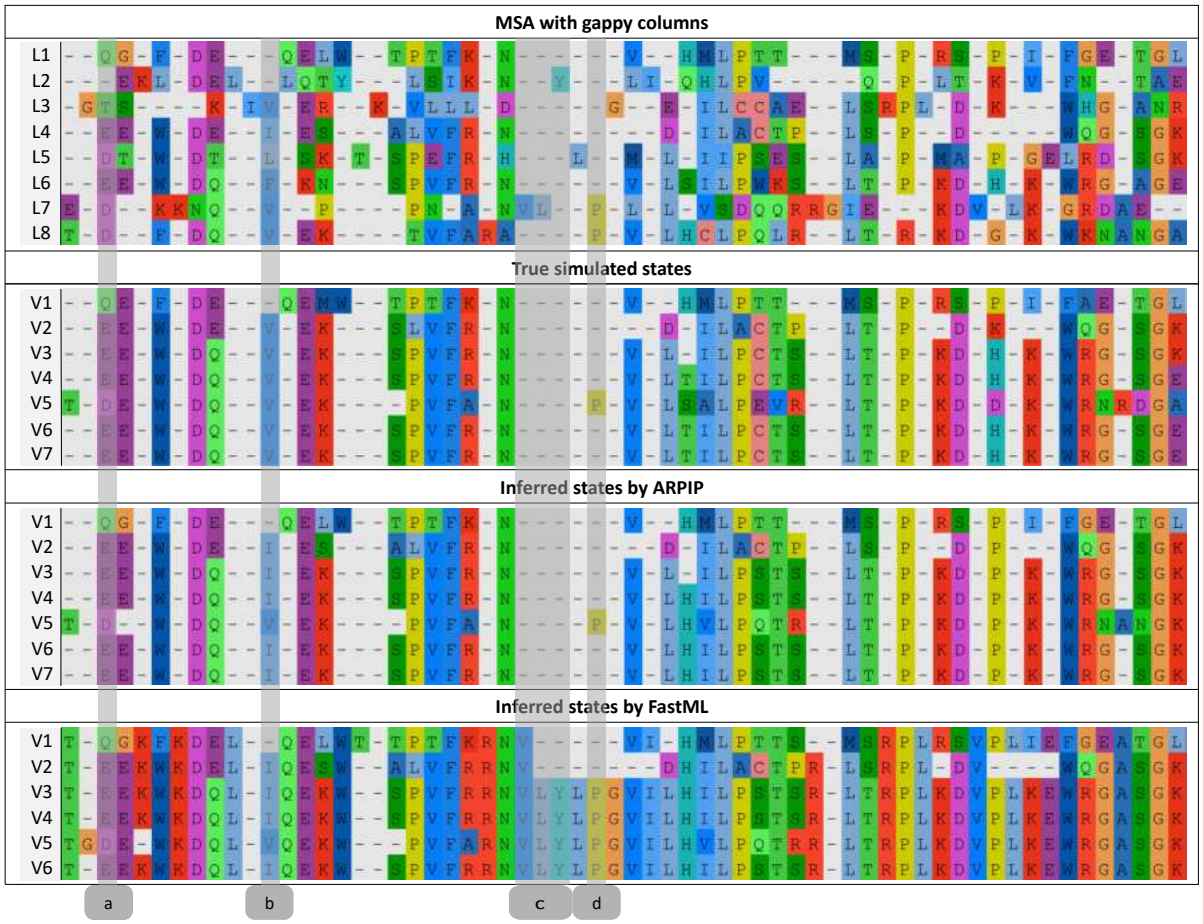


Figure 3.6: A snippet from the PIP simulated dataset containing the true simulated MSA and ancestors and the ancestral sequences predicted by ARPIP and FastML. a) A region where both ARPIP and FastML accurately inferred the ancestral states. b) A region where both algorithms estimated the ancestral character incorrectly. c) A region where FastML inferred ancestral characters even though there were none in the simulation. d) A region where there was a single ancestral character but FastML inferred its position incorrectly. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

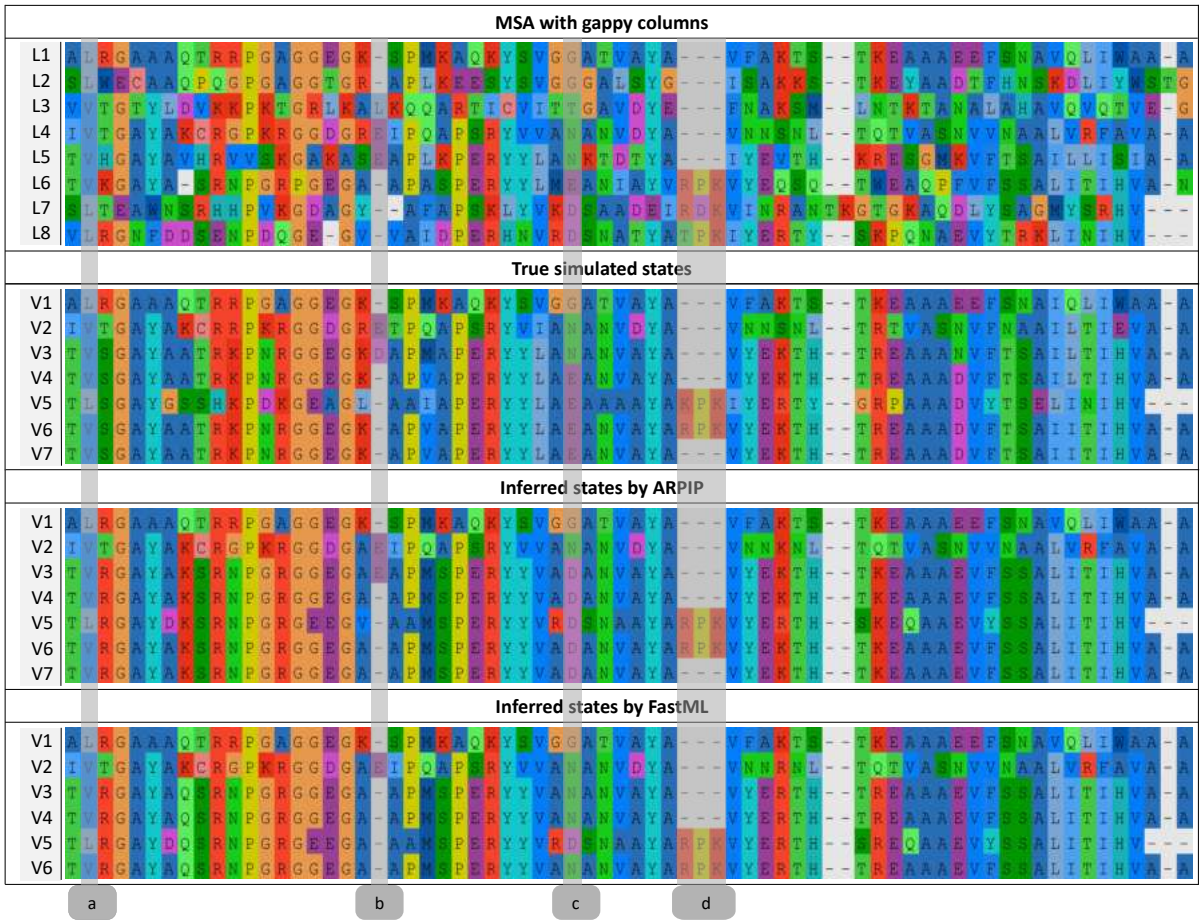


Figure 3.7: A snippet from the INDELible simulated dataset with indel rate 0.01 containing the true simulated MSA, ancestors and the ancestral sequences predicted by ARPIP and FastML. a) A region where both ARPIP and FastML accurately inferred the ancestral states. b) A region that the FastML inferred the gaps incomplete while ARPIP missed the the character state. c) A region where both algorithms estimated most of the ancestral character incorrectly. d) A region where both methods inferred the ancestral states including gaps positions correctly. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

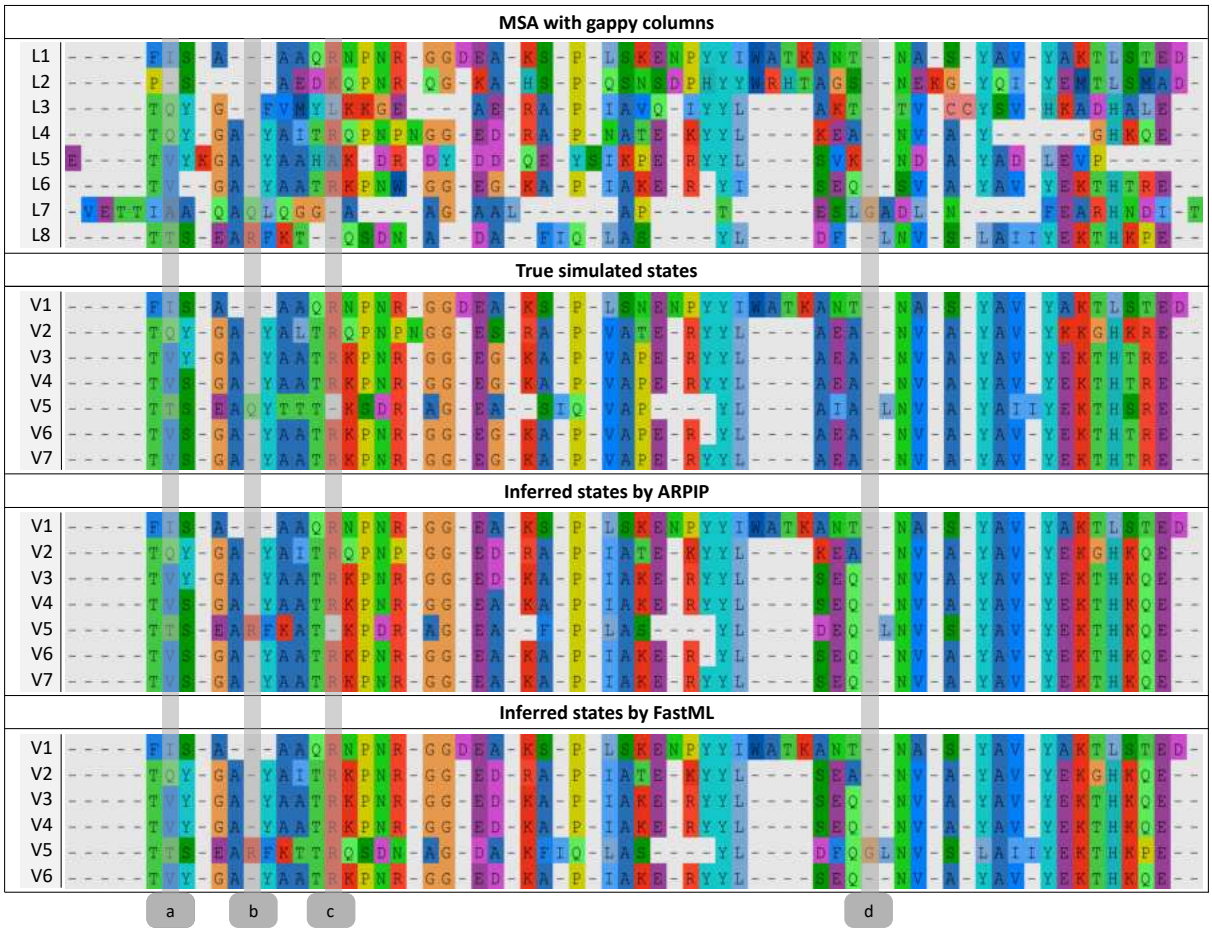


Figure 3.8: A snippet from the INDELible simulated dataset with indel rate 0.05 containing the true simulated MSA and ancestors and the ancestral sequences predicted by ARPIP and FastML. a) A region where both ARPIP and FastML accurately inferred the ancestral states. b) A region where both algorithms estimated the indel events correctly but the ancestral character incorrectly. c) A region where FastML missed the gap character but ARPIP inferred it correctly. d) A region where FastML inferred ancestral characters even though there were none in the simulation. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

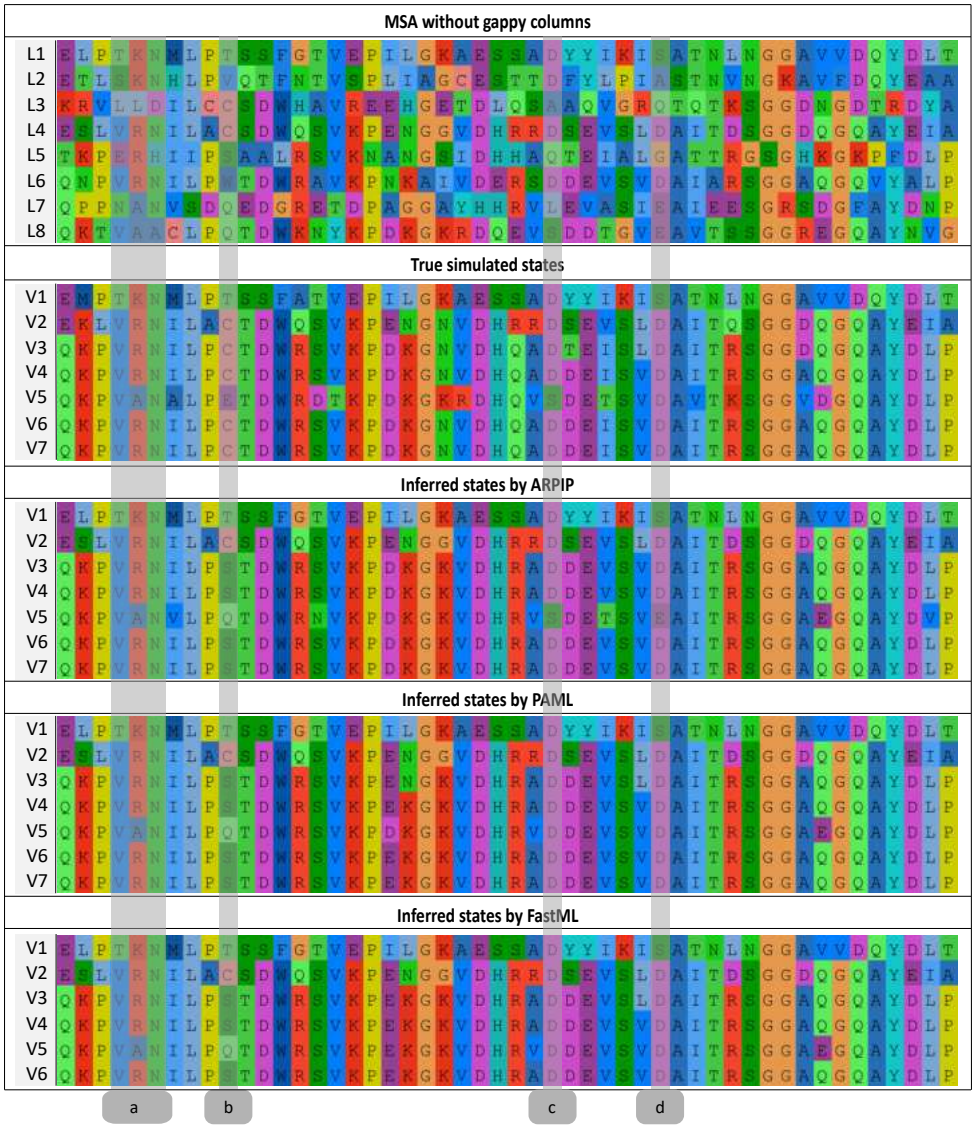


Figure 3.9: A gapless snippet from the PIP simulated dataset containing the true simulated MSA and ancestors and the ancestral sequences predicted by ARPIP, PAML and FastML. a) A region where all algorithms accurately inferred the ancestral state. b) A region where all algorithms made mistakes. c) A region where FastML and PAML made incorrect inferences but ARPIP inferred the ancestral state correctly. d) A region where all algorithms accurately inferred the ancestral states except ARPIP. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

inferences with those by PAML and FastML, showing that our approach performs just as well for data without gaps.

While indel events represent a major mutational process of gene evolution [Söding and Lupas 2003], they are rarely accounted for in ASR. ARPIP expands the reconstruction possibilities to include more divergent and gappy sequences allowing us to study a wider range of resurrected ancestral molecules, investigating the functional importance of indels in ancestral proteins. This is particularly valuable for proteins separated by large divergences or within "more flexible" loop regions, as indels frequently occur in regions where amino acid sequences are not well conserved [Taylor et al. 2004].

While single residue indel modelling may be viewed as a limitation, certain types of genetic material exhibit specifically these kind of indel events more often than others. For example, single nucleotide indels are predominant between recently diverged DNA sequences from various organisms [Tao et al. 2007] and in non-coding DNA sequences [Yamane et al. 2006]. While most ASR is done on coding sequences to investigate the properties of reconstructed proteins, it has recently been shown that many trait-associated loci, including some associated with disease, lie outside protein-coding regions [Kellis et al. 2014]. ARPIP can be used to reconstruct non-coding sequences with meaningful biological assumptions, which could be an additional avenue of exploration for disease-related ASR.

ARPIP paves the way for even more new types of indel analyses. The approach can be expanded to analyse the patterns of insertions and deletions by including rate heterogeneity, for example, allowing us to detect lineage-specific patterns through time. We can include site-specific indel rate variation, allowing us to see the difference in indel evolution in different functional regions of proteins such as loops or active sites. Then, we can investigate the occurrence and consequences of indels in specific regions such as indel-tolerant regions of the genome and relation between gene function and indel frequency [Taylor et al. 2004]. Moreover, in the long run our method can be used to extend and potentially improve more sophisticated probabilistic approaches such as [Groussin et al. 2014], which accounts not only for gene-trees but also for species history, therefore including gene gain/loss and horizontal transfer in the inference.

While some other methods have attempted to reconstruct exact indel histories, the only other currently existing method in the frequentist framework can only handle small datasets [Diallo et al. 2007]. Even though PIP makes simplifying assumptions like site independence, which only allows us to model single residue indels, the explicit evolutionary model makes indel events interpretable. Moreover, as the method has linear time complexity, we can use this approach as a building block in integrated alignment-tree-ancestor inference [Pečerska et al. 2021], using the indel points under PIP as a starting point for integrating more complex models of indel

evolution, e.g. moving on to long indel models.

Like other likelihood based approaches, our method in theory allows us to explore both optimal and suboptimal reconstructions in follow-up analyses. It has been argued that a single reconstruction (i.e. a point estimate) can be inadequate in cases when the likelihood surface is nonconvex and contains multiple local optima [Joy et al. 2016], which can lead to systematic bias [Yang 2014, p.131]. Since ARPIP is in essence an empirical Bayes method, we can extend the method to account for the uncertainty in our estimates by working with probability profiles of characters and gaps rather than inferences fixed to the optimal estimates [Williams et al. 2006].

3.4 Availability of the experimental data and code

The proposed algorithm has been implemented based on Bio++ open source library [Guéguen et al. 2013] using C++ programming language. Our code and experimental data used in this manuscript is freely available at `https://github.com/acg-team/bpp-ARPIP` under GNU GPLv3 licence. One can find the data and a brief user manual on our GitHub page.

3.5 Funding

This work was supported by the Swiss National Science Foundation (SNF) grant 31003*A*_176316 to M.A. The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

3.6 Acknowledgements

We would like to thank A. Bouchard-Côté (University of British Columbia) for providing his code JavaPIP to simulate sequences under the PIP, and our master student J. Peechatt for his preliminary work which helped to develop this method.

3.7 Appendices

3.7.1 Appendix 1

3.7.1.1 The PIP description

3.7.1.1.1 Insertion probability For every node $v \in V$, the probability of inserting a single character on edge $e = (v \to pa(v))$ is proportional to the branch length $b(v)$:

$$\iota(v) = \frac{1}{||\tau|| + \mu^{-1}} \cdot \begin{cases} b(v) & \text{if } v \neq \Omega \\ \mu^{-1} & \text{if } v = \Omega, \end{cases} \tag{3.1}$$

where $\|\tau\|$ is the sum of all branch lengths (total tree length).

3.7.1.1.2 Survival probability The survival probability $\beta(v)$ for a character inserted along edge e is given by:

$$\beta(v) = \begin{cases} \left(1-\exp(-\mu b(v))\right)/(\mu b(v)) & \text{if } v \neq \Omega \\ 1 & \text{if } v = \Omega, \end{cases} \tag{3.2}$$

3.7.1.1.3 Pure survival probability To compute the probability that a character inserted before edge e survives along edge e, we define the pure survival probability $\zeta = \exp(-\mu b(v))$ associated with node v. Likewise, the probability that the character will not survive along the edge is $1 - \zeta(v)$. The pure survival probability $\zeta(v)$ differs from the definition of $\beta(v)$ with regard to the insertion location. While in $\zeta(v)$ the character is already present at $pa(v)$, in $\beta(v)$ the character is inserted in a random location along the edge associated with v. In both cases, however, the character is required to survive until node v [Maiolo 2019].

3.7.1.1.4 MSA column probability For each individual MSA column m, the probability $p(m)$ is defined by marginalizing over all possible homology paths that may have created that MSA column:

$$p(m) = \sum_{v \subset V} \iota(v) f_v, \tag{3.3}$$

where f_v denotes the probability of all possible homology paths for MSA column m and the subtree rooted at v. Note that f_v can be zero for some nodes, namely the ones where an insertion

is impossible given the data in m.

3.7.1.2 The probability of all possible homology paths.

$$f_v = \begin{cases} \mathbb{1}[v \in \mathscr{A}]\beta(v)\sum_{\sigma\in\Sigma}\pi_\varepsilon(\sigma)\tilde{f}_v(\sigma) & \text{if } m \neq m_\emptyset \\ 1-\beta(v)+\beta(v)\sum_{\sigma\in\Sigma}\pi_\varepsilon(\sigma)\tilde{f}_v(\sigma) & \text{o.w.,} \end{cases} \tag{3.4}$$

where $m_\emptyset$ is MSA column full of gaps and

$$\tilde{f}_v(\sigma) = \begin{cases} \mathbb{1}[m_v = \sigma] & \text{if } v \in \mathscr{L} \\ \prod_{w\in\text{child}(v)}\left[\sum_{\sigma'\in\Sigma_\varepsilon}\exp(b(w)Q_\varepsilon)_{(\sigma,\sigma')}\tilde{f}_w(\sigma')\right] & \text{o.w.,} \end{cases} \tag{3.5}$$

and $\mathbb{1}[.]$ is the indicator function. The recursive function $\tilde{f}_v(\sigma)$ denotes the partial likelihood of a single character under the substitution-deletion events.

3.7.2 Appendix 2

3.7.2.1 Detailed description of IndelPoints algorithm

The goal of the IndelPoints algorithm is to find the combination of insertion and deletion points with the highest conditional probability for a given site. These indel points define a homology path which is the best possible explanation of the homologous characters in the alignment. We infer these indel points using a greedy progressive approach, finding the homology path with the highest conditional probability in each node of the tree and progressively extending the path at other nodes. This way, for a given site we find the best partial homology path at each tree node, and can select the best homology path on the whole tree once we reach the root.

The space of possible homology paths is constrained by the definition of PIP, meaning that per site there can only be a single insertion location and multiple deletions. Moreover, the sets of nodes that are possible insertion and deletion locations are non-overlapping. A homology path will have a non-zero conditional probability if and only if it explains the full site in the alignment, i.e. it contains one of the nodes $v \in \mathscr{A}$ as an insertion location. This means that in any node that is not a possible insertion location the conditional probability can always be set to 0.

Based on the properties of PIP, we can describe the evaluation separately for possible deletion nodes and for all other nodes in the tree. The set $\mathscr{G}$ of potential deletion locations defines the root nodes of subtrees that went extinct at the leaves. For simplicity we will call these nodes

extinction nodes. For each of these nodes, we will find the set of deletion nodes with the highest conditional probability and will set the conditional probability of the full homology path to 0. We will call the complementary set of nodes $v \notin \mathscr{G}$ the set of survival nodes. In these nodes, the character either got inserted along the branch $\text{pa}(v) \to v$ for $v \in \mathscr{A}$, or got inserted in an ancestor of v for $v \notin \mathscr{A}$. In either case the character definitely survived until v.

We traverse the tree in post order, evaluating possible homology paths per tree node. The algorithm is described separately for the two node sets (extinction nodes and survival nodes) however all necessary computation can be done in a single tree traversal. It is important to note that here the values are computed at the highest point of the branch $\text{pa}(v) \to v$ rather than at the bottom of the branch leading to v. To simplify notation, we will use v_L and v_R to denote the left and right child node of v respectively.

3.7.2.1.1 Evaluating extinction nodes For each extinction node $v \in \mathscr{G}$ we set the insertion node to none ($\mathscr{I}_v = \emptyset$) and consequently the probability of the best global homology path to 0 ($p_v = 0$).

We need to learn whether the deletion was more likely along the branch leading to v or more likely in the child subtrees. The character could have gone extinct before reaching node v, which is the probability of deletion happening along the branch $\text{pa}(v) \to v$. Moreover, if v is a leaf ($v \in \mathscr{L}$), we can certainly state that it is the current best deletion node ($\mathscr{D}_v = v$) and that f_v at that node represents the certain deletion ($f_v = 1 - \zeta(v)$, extinction probability). On the other hand, if the node v is not a leaf node ($v \notin \mathscr{L}$), we compare the deletion probability at $\text{pa}(v) \to v$, $1 - \zeta(v)$, with the product of survival along $\text{pa}(v) \to v$ and subsequent deletion at or below nodes v_R and v_L, $f_{v_\text{R}} f_{v_\text{L}} \zeta(v)$. Among the two, we select the value that represents the best scenario and set f_v and $\mathscr{D}_v$ appropriately ($f_v = 1 - \zeta(v)$ and $\mathscr{D}_v = \{v\}$ or $f_v = f_{v_\text{R}} f_{v_\text{L}} \zeta(v)$ and $\mathscr{D}_v = \{\mathscr{D}_{v_\text{R}} \cup \mathscr{D}_{v_\text{L}}\}$ respectively).

The pseudocode for extinction node evaluation is shown in Algorithm 1, and the corresponding flowchart is available in the supplemental materials (see Fig. 3.10).

3.7.2.1.2 Evaluating survival nodes All nodes $v \notin \mathscr{G}$ are definite survival nodes – meaning that the character survived along the branch $\text{pa}(v) \to v$. We distinguish two types of survival nodes, $v \in \mathscr{A}$ (potential insertion nodes) and $v \notin \mathscr{A}$, nodes where an insertion could not have happened. $v \notin \mathscr{A}$ is the simpler one, as the homology path at such nodes will not have an associated insertion node, while a full homology path needs to have an insertion. This means that for all such nodes the probability of a homology path will be 0 ($p_v = 0$) and the set of insertion nodes will be empty ($\mathscr{I}_v = \emptyset$). In case $v \in \mathscr{A}$, we compute the non-zero homology path probability according to the formulas defined for PIP, $p_v = \iota(v)\beta(v)$ for a leaf node and

Algorithm 1: Evaluating extinction nodes.

Input: Phylogenetic tree τ

Output: Homology path with the highest probability $\mathscr{H}(m)$

for $\forall v \in \mathscr{G}$ **do**
  $p_v = 0$;
  if $v \in \mathscr{L}$ **then**
    $f_v = 1 - \zeta(v)$;
    $\mathscr{H}_v = \{\mathscr{I}_v = \emptyset, \mathscr{D}_v = \{v\}\}$;
  else
    if $\zeta(v) f_{v_\mathrm{R}} f_{v_\mathrm{R}} < 1 - \zeta(v)$ **then**
      $f_v = 1 - \zeta(v)$;
      $\mathscr{H}_v = \{\mathscr{I}_v = \emptyset, \mathscr{D}_v = \{v\}\}$;
    else
      $f_v = \zeta(v) f_{v_\mathrm{R}} f_{v_\mathrm{L}}$;
      $\mathscr{H}_v = \{\mathscr{I}_v = \emptyset, \mathscr{D}_v = \mathscr{D}_{v_\mathrm{R}} \cup \mathscr{D}_{v_\mathrm{L}}\}$;
    end
  end
end

$p_v = \iota(v)\beta(v) f_{v_\mathrm{R}} f_{v_\mathrm{L}}$ for an internal node. We also set $\mathscr{I}_v = v$, representing a possible insertion at this node. While the set $\mathscr{A}$ may contain multiple nodes, an instance of a homology path can only have a single insertion location. This implies that when we compute the probability of a homology path for a node $v \in \mathscr{A}$, we essentially assume that it is the only possible insertion location while all other nodes are treated as regular nodes in the tree.

The conditional probability of survival is independent of whether v is a potential insertion node or not. If the node is a leaf ($v \in \mathscr{L}$), the character will have certainly survived, thus $f_v = \zeta(v)$, and the deletion node set is empty $\mathscr{D}_v = \emptyset$. On the other hand, if v is an internal node, we need to account for the survival along branch $\mathrm{pa}(v) \to v$ and propagate any possible deletion nodes that were already selected, thus $f_v = f_{v_\mathrm{R}} f_{v_\mathrm{L}} \zeta(v)$ and $\mathscr{D}_v = \mathscr{D}_{v_\mathrm{R}} \cup \mathscr{D}_{v_\mathrm{L}}$.

Reaching the root of the tree Ω, we will have processed all the nodes in the tree and computed the probabilities of the homology paths conditioned on the insertion point at each node. These probabilities will only be non-zero for nodes $v \in \mathscr{A}$, which allows us to choose the best homology path simply as $\mathscr{H}_{\mathrm{argmax}(p_v)}$ – the homology path corresponding to the node with the highest probability. We then extract the subtree τ_m rooted at $\mathscr{I}_{\arg\max(p_v)}$ and pruned at $\mathscr{D}_{\arg\max(p_v)}$ from the tree τ.

The pseudocode for survival node evaluation is presented in Algorithm 2 and the corresponding flowchart is available in the supplemental materials (see Fig. 3.11).

Algorithm 2: Evaluating survival nodes

Input: Phylogenetic tree τ

Output: Homology path with the highest conditional probability $\mathscr{H}(m)$

for $\forall v \notin \mathscr{G}$ **do**
  if $v \in \mathscr{L}$ **then**
    $p_v = \mathbb{1}[v \in \mathscr{A}]\iota(v)\beta(v)$;
    $f_v = \zeta(v)$;
    $\mathscr{H}_v = \{\mathscr{I}_v = \mathbb{1}[v \in \mathscr{A}]v, \mathscr{D}_v = \emptyset\}$;
  else
    $p_v = \mathbb{1}[v \in \mathscr{A}]\iota(v)\beta(v)f_{v_\mathrm{R}}f_{v_\mathrm{L}}$;
    $f_v = \zeta(v)f_{v_\mathrm{R}}f_{v_\mathrm{L}}$;
    $\mathscr{H}_v = \{\mathscr{I}_v = \mathbb{1}[v \in \mathscr{A}]v, \mathscr{D}_v = \mathscr{D}_{v_\mathrm{R}} \cup \mathscr{D}_{v_\mathrm{L}}\}$;
  end
end
return $\mathscr{H}_{\arg\max(p_v)}$;

3.7.3 Appendix 3

3.7.3.1 Detailed DP joint ASR under PIP

Similar to how it is done in FastML [Pupko et al. 2000], we reconstruct the ancestral characters in two consecutive steps. First, we compute the likelihoods using dynamic programming (DP), and then use the precomputed values for joint ASR. However, instead of reconstructing the ancestral states on the whole tree τ, we work on the subtree τ_m extracted using the most likely indel history for the site m under PIP model.

3.7.3.1.1 DP likelihood computation In the forward phase, as in Pupko's ASR method, we traverse the tree τ_m in post-order. While visiting each internal node v, we compute the likelihood $\mathrm{Lk}_v(i)$ and the corresponding best ancestral character state $\mathrm{CS}_v(i)$ for each character i in the alphabet ($i \in \Sigma$). We assume that the best ASR in a subtree rooted at the node v is independent of the rest of the tree. $\mathrm{Lk}_v(i)$ is the likelihood of the best reconstruction on this subtree conditioned on the parent of node v has the character i at that site. And $\mathrm{CS}_v(i)$ is the character state assigned to node v in this optimal conditional reconstruction.

To compute the likelihood values and the corresponding candidate character states, we have to determine whether a node is a leaf or not. Let j be the observed character at v. If $v \in \mathscr{L}$, we assign $\mathrm{CS}_v(i) = j$ and the likelihood value is $\mathrm{Lk}_v(i) = \mathrm{P}_{ij}(v)$, where $\mathrm{P}(v) = \exp(\mathrm{b}(v) \cdot \mathrm{Q}_\varepsilon)$ is the transition probability matrix along the branch $\mathrm{pa}(v) \to v$. If a node v is not a leaf ($v \notin \mathscr{L}$),

we compute the likelihood value and the candidate characters for all the non-root internal nodes based on the values already computed in its children. For each $i \in \Sigma$ the likelihood value is $\mathrm{Lk}_v(i) = \max_j(\mathrm{P}_{ij}(v) \cdot \mathrm{Lk}_{v_\mathrm{L}}(j) \cdot \mathrm{Lk}_{v_\mathrm{R}}(j))$ and the candidate character is the value of j producing the previous maximization $\mathrm{CS}_v(i) = \arg\max_j(\mathrm{P}_{ij}(v) \cdot \mathrm{Lk}_{v_\mathrm{L}}(j) \cdot \mathrm{Lk}_{v_\mathrm{R}}(j))$. We compute there values until all the nodes have been visited except for the root. While visiting the root, the likelihood is computed slightly differently as the transition probabilities are replaced with the alphabet equilibrium frequencies $\pi_{\varepsilon i}$, $\mathrm{Lk}_v(i) = \pi_{\varepsilon i} \cdot \mathrm{Lk}_{v_\mathrm{L}}(i) \cdot \mathrm{Lk}_{v_\mathrm{R}}(i)$. Since the root has no possible ancestors, we define the single candidate character state as $\mathrm{CS}_\Omega = \arg\max_i(\mathrm{Lk}_\Omega)$. The pseudocode of the algorithm is shown in Algorithm 3.

Algorithm 3: Computing the likelihood

Input: Phylogenetic subtree τ_m
MSA column m
Extended substitution model Q_ε
Extended equilibrium frequencies π_ε

Output: Likelihood array of node v Lk_v
Candidate character array of node v with ML values CS_v

Ω_{τ_m} = root of subtree τ_m;
$\mathrm{P}(v) = \exp(\mathrm{b}(v) \cdot \mathrm{Q}_\varepsilon)$;

if $v \in \mathscr{L}$ **then**
    $j =$ observed character at v;
    for $\forall i \in \Sigma$ **do**
        $\mathrm{CS}_v(i) = j$;
        $\mathrm{Lk}_v(i) = \mathrm{P}_{ij}(v)$;
    end
else
    v_L = left child node;
    v_R = right child node;
    if *v is equal to* Ω_{τ_m} **then**
        for $\forall i \in \Sigma$ **do**
            $\mathrm{Lk}_v(i) = \pi_{\varepsilon i} \cdot \mathrm{Lk}_{v_\mathrm{L}}(i) \cdot \mathrm{Lk}_{v_\mathrm{R}}(i)$;
        end
        $\mathrm{CS}_\Omega = \arg\max_i(\mathrm{Lk}_v)$;
    else
        for $i \in \Sigma$ **do**
            for $j \in \Sigma$ **do**
                $S(j)$[1] $= \mathrm{P}_{ij}(v) \cdot \mathrm{Lk}_{v_\mathrm{L}}(j) \cdot \mathrm{Lk}_{v_\mathrm{R}}(j)$;
            end
            $\mathrm{Lk}_v(i) = \max(S)$;
            $\mathrm{CS}_v(i) = \arg\max_j(S)$;
        end
    end
end

[1] S stands for Single Character State selected in this step of the algorithm.

3.7.3.1.2 Joint ASR After computing the likelihoods of all potential ancestral characters, we traverse the tree in pre-order from the root of the tree τ_m to select the single most likely ancestral character per node A_v. We assign the most likely character state at the root by picking the character i that maximizes the likelihood value $CS_{\Omega_{\tau_m}} = \arg\max_i(Lk_v)$. For all other internal nodes in the traversal we assign $CS_v(A_{pa(v)})$ as the best reconstruction (for example if the state reconstructed for the parent of v is proline $(i = P)$, the ancestral state of v will be $CS_v(i = P)$). We continue the process until the most likely character is assigned at all the internal nodes. The formula is presented below where Ω_{τ_m} is the root of subtree τ_m:

$$A_v = \begin{cases} CS_v(\arg\max(Lk_v)) & v = \Omega_{\tau_m} \\ CS_v(A_{pa(v)}) & \text{o.w.} \end{cases} \tag{3.6}$$

3.8 Supplemental materials

Data available from the Dryad Digital Repository: `https://doi.org/10.5061/dryad.wstqjq2nj`.

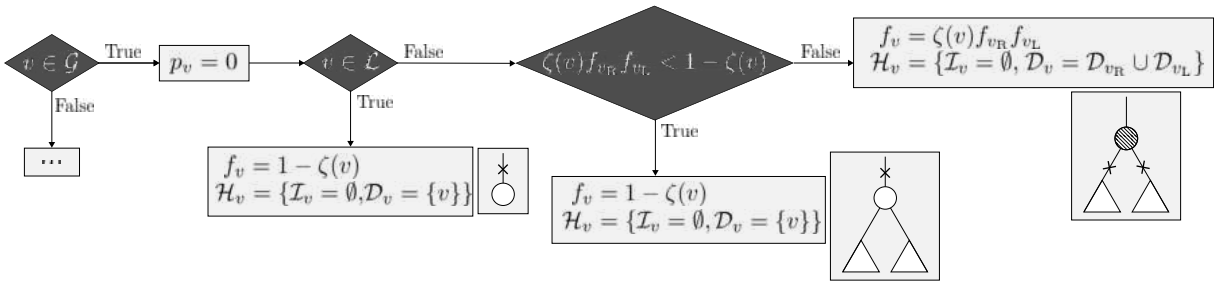


Figure 3.10: Evaluating extinction nodes, where $\mathscr{H}_v$ defines the homology path of node v, f_v is the probability of that homology path and p_v is the probability of the best homology path. $\mathscr{I}_v$ is the set of possible insertion points while $\mathscr{D}_v$ is the set of all possible deletion points for node v. $\zeta(v)$ denotes the pure survival probability of node v. Moreover, $\mathscr{G}$ and $\mathscr{L}$ represent the set of potential deletion points and leaves, respectively.

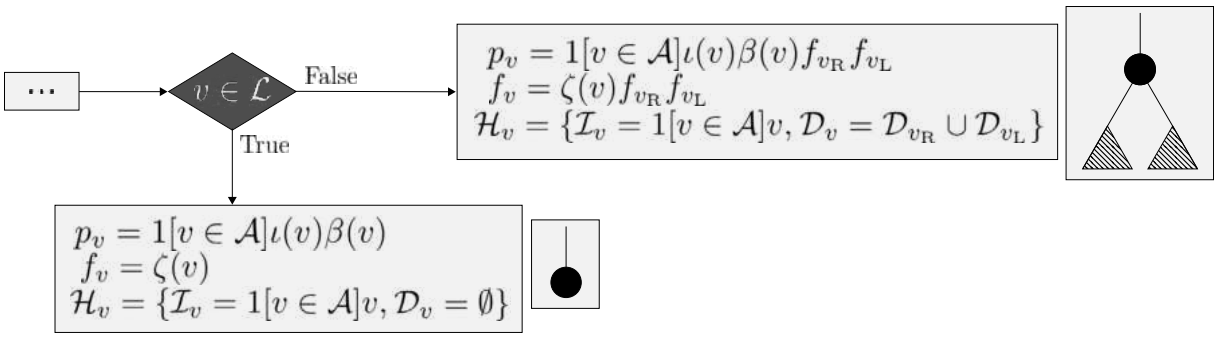


Figure 3.11: Evaluating survival nodes. $\mathscr{H}_v$ defines the homology path of node v represents by the set of possible insertion points $\mathscr{I}_v$ and the set of all possible deletion points $\mathscr{D}_v$. Moreover, f_v is the probability of that homology path while p_v is the probability of the best homology path. $\iota(v)$, $\beta(v)$ and $\zeta(v)$ are respectively insertion, survival and pure survival probabilities at node v. $\mathscr{A}$ represents the set of potential insertion points and $\mathscr{L}$ shows the set of leaves.

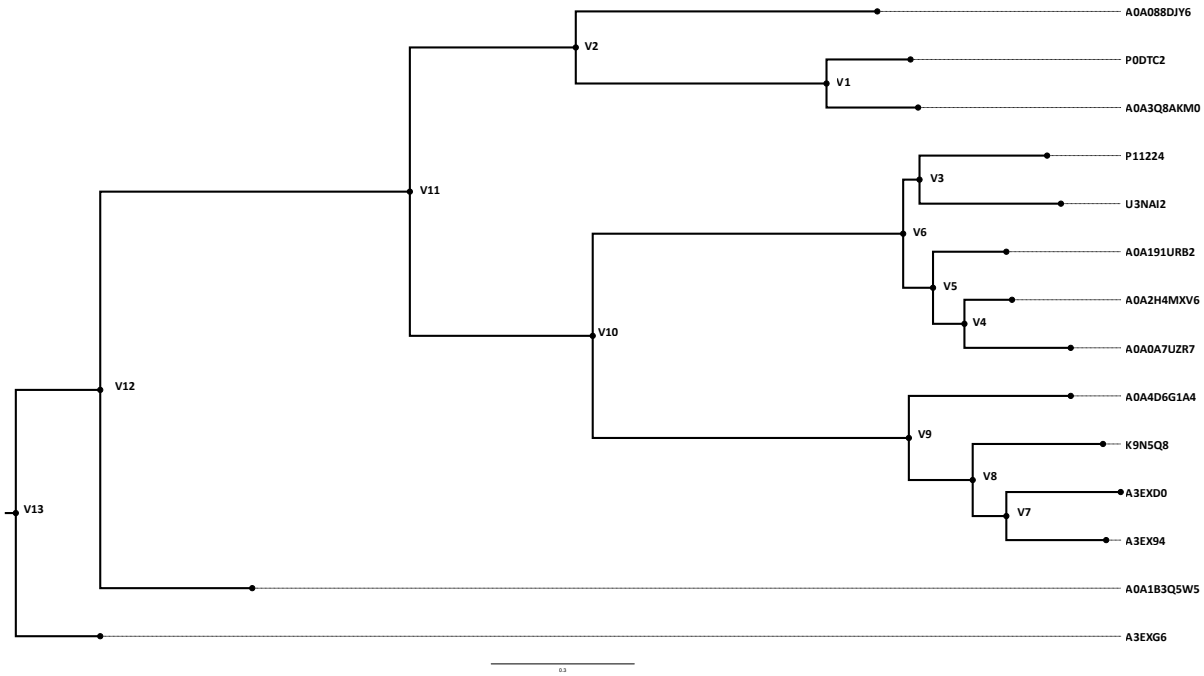


Figure 3.12: Illustration of the rooted *Betacoronavirus* phylogenetic tree reconstructed by PhyML 3.0 from the PRANK alignment. Note that the original tree was unrooted which ARPIP used mid-point rooting method to make the tree rooted.

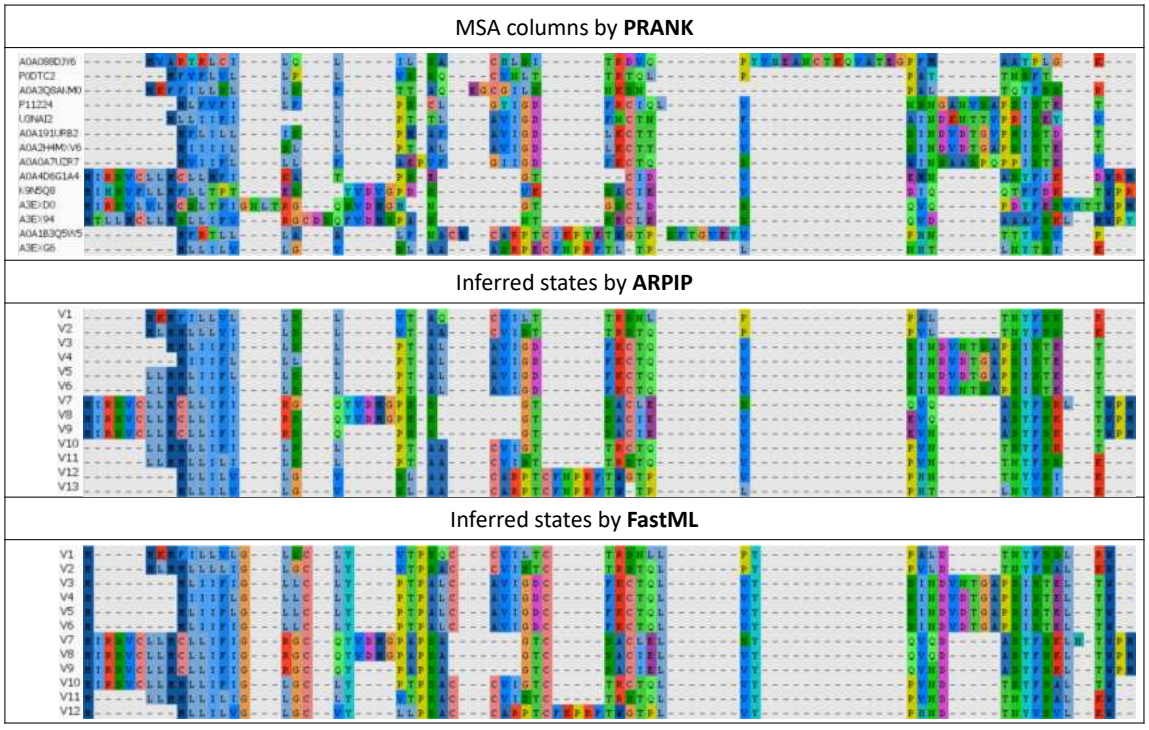


Figure 3.13: A snippet from the *CoV* dataset containing the MSA inferred by PRANK and the ancestor sequences predicted by ARPIP and FastML. The tree is obtained by PhyML 3.0, shown in Figure 3.12. Note that FastML algorithm works on an unrooted tree which, compared to ARPIP, resulted in one fewer internal sequence reconstructed (due to the absence of the root node).

Chapter 4

Single-character insertion-deletion model preserves long indels in ancestral sequence reconstruction

This chapter presents a manuscript draft that was uploaded on bioRxiv and submitted to an international peer-review for publication. This chapter includes the empirical data analysis investigating the dynamics of insertions and deletions in mammalian orthologous proteins and the examination of the ancestral reconstruction of multiple-character indels under the PIP. The experimental data used in this manuscript is freely available from `https://github.com/acg-team/single-char-indel-ASR-preserves-long-indels`. The manuscript, where I am the first author, is publicly available as bioRxiv at `https://doi.org/10.1101/2024.03.09.584071`.

Single-character insertion-deletion model preserves long indels in ancestral sequence reconstruction

Gholamhossein Jowkar[1,2,3,*], Jūlija Pečerska[2,3], Manuel Gil [2,3], and Maria Anisimova [2,3]

[1] *University of Neuchâtel, Institute of Biology, CH-2000 Neuchâtel, Switzerland*

[2] *Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*

[3] *Zurich University of Applied Sciences, LSFM, ICLS, CH-8820, Wädenswil, Switzerland*

**Gholamhossein Jowkar, ZHAW, School of Life Sciences and Facility Management, Applied Computational Genomics Group, Schloss 1, 8820 Wädenswil, Switzerland, E-mail: jowk@zhaw.ch*

Abstract

Insertions and deletions (indels) play a significant role in genome evolution across species. Realistic modelling of indel evolution is challenging and is still an open research question. Several attempts have been made to explicitly model multi-character (long) indels, such as TKF92, by relaxing the site independence assumption and introducing fragments. However, these methods are computationally expensive.

On the other hand, the Poisson Indel Process (PIP) assumes site independence but allows one to infer single-character indels on the phylogenetic tree, distinguishing insertions from deletions. PIP's marginal likelihood computation has linear time complexity, enabling ancestral sequence reconstruction (ASR) with indels in linear time. Recently, we developed ARPIP, an ASR method using PIP, capable of inferring indel events with explicit evolutionary interpretations.

Here, we investigate the effect of the single-character indel assumption on reconstructed ancestral sequences on mammalian protein orthologs and on simulated data. We show that ARPIP's ancestral estimates preserve the gap length distribution observed in the input alignment. In mammalian proteins the lengths of inserted segments appear to be substantially longer compared to deleted segments. Further, we confirm the well-established deletion bias observed in real data.

To date, ARPIP is the only ancestral reconstruction method that explicitly models insertion and deletion events over time. Given a good quality input alignment, it can capture ancestral long indel events on the phylogeny.

Keywords: ancestral sequence reconstruction, insertion, deletion, mammalian genomics, Poisson indel process, indel pattern, long indel, gap length distribution.

4.1 Introduction

Insertion and deletion (indel) events produce significant amounts of natural variation in species genomes. Consequently, indels make a major contribution to complex evolutionary processes. Today indel variants in genomic sequences can be reliably documented and studied due to improvements in sequencing methods. In closely related species, differences attributed to indels (per base pair) are several-fold more frequent than substitution events [Britten et al. 2003, Wetterbom et al. 2006]. In the human genome, up to a quarter of all genomic variants are due to indels, most of which are very short [Mills et al. 2006]. While indels are distributed across both coding and non-coding parts of genomes, they are far more frequent in non-coding sequences. Compared to substitutions, indel changes are expected to have a stronger deleterious effect on functional proteins [Tóth-Petróczy and Tawfik 2013], also explaining their lower prevalence in coding sequences. Despite this, many deleterious coding indel variants persist in the human population and can cause disease-related gene defects (e.g., [Chuzhanova et al. 2003]).

In comparative studies of sequence evolution, indels are represented as gaps in alignments of homologous sequences. With growing divergence, different indel events can merge and overlap, masking the mutational history. Nevertheless, alignment gaps carry much phylogenetic information [Dessimoz and Gil 2010], which can provide valuable insights for evolutionary studies when analyzed correctly. However, properly modelling the evolutionary process of insertions and deletions is challenging from the computational and modelling perspective, and there is no gold standard in the field. In fact, many evolutionary studies either completely ignore indels, or heavily trim indel-rich sequence regions, due to the lack of software tools implementing appropriate models. Disentangling individual insertion and deletion events based on the observed gap distributions in a multiple sequence alignment (MSA) requires stochastic models of sequence evolution that also include the insertion and deletion processes over time. Substitutions are traditionally described via Markov models, which assume site independence, while indels violate this assumption since each indel event can involve multiple residues. Therefore, models properly including these events tend to be computationally expensive.

The first evolutionary model with indels, TKF91, lifted the assumption of site independence and described single-character indels via a birth-death process [Thorne et al. 1991]. As TKF91 models single-character events, it implies a linear gap cost in the MSA inference, but due to the non-independence of sites, the complexity of computing the marginal likelihood under this model is exponential in the number of taxa. Bouchard-Côté and Jordan proposed the PIP model, a close relative of TKF91, where insertions follow the Poisson process while deletions are added to the Markov substitution model as an absorbing state. The complexity of marginal likelihood computation under the PIP model is reduced to linear, which allows for this model to be adopted

for phylogenetic inferences [Zhai and Bouchard-Côté 2017, Maiolo et al. 2018; 2021, Jowkar et al. 2023]. However, like TKF91, PIP explicitly models only single-character indels.

Modelling longer indels as several independent single-character events lacks biological realism and could lead to biases such as homology histories with too many events, alignments with scattered gaps, and high indel rates. Some evolutionary indel models allow long indels [Thorne et al. 1992, Miklós et al. 2004, De Maio 2021]. For example, the TKF92 model, an extension of TKF91, is also a birth-death process but with indels happening as unbreakable multiple-site fragments with a geometric length distribution [Thorne et al. 1992]. This modelling assumption, however, means that TKF92 cannot explain overlapping indels. The "long indel" model [Miklós et al. 2004] relaxed the unbreakable fragment assumption but assumed infinite sequences. Both these models can be considered an approximation of the Generalised Geometric Indel (GGI) model [Holmes 2020]. However, while the lengths of individual indels have a geometrical distribution, the length distribution of observed gaps in the alignment is not geometric in general. Considering that models with long indels also tend to be computationally slow, these are currently of little practical value for large datasets.

Computationally, PIP holds promise for practical phylogenetic analyses despite the single-character indel assumption. For example, we showed that PIP-based alignment inference can pick up multiple-character indels (long indels) when the data strongly suggests this [Maiolo et al. 2018; 2021]. Zhai and Bouchard-Côté demonstrated that modelling indel evolution and indel rate variation improves the accuracy of phylogeny reconstruction when using the PIP model and its generalizations.

Recently, we proposed a PIP-based ancestral sequence reconstruction (ASR) approach implemented in ARPIP [Jowkar et al. 2023]. Apart from Bayesian MCMC implementations (e.g., Historian [Holmes 2017]), ARPIP is the only ASR method that uses an explicit model of indel evolution and can infer the specific locations of insertions and deletions on the tree. Another popular ASR method is FastML-webserver [Ashkenazy et al. 2012], which uses the so-called "indel-coding" method to include indels. This approach does not include a proper statistical model of insertion and deletion and implies that a deleted character can be reinserted. GRASP [Ross et al. 2022], another recent method, accommodates indels in the ASR inference by representing sequences as partial order graphs. However, as with indel-coding, deleted characters can be reinserted, and there is no explicit model governing the indel process.

4.2 The goals of this study

Having an explicit model of indel evolution is desirable; however, an over-simplistic model could also have a detrimental effect on the resulting inferences, including overestimation of indel rates and scattered ancestral sequence alignments by including too many single-character gaps. Therefore, we aim to investigate whether using the single-character indel assumption negatively impacts ASR. Since ASR methods typically take a fixed MSA and phylogeny as input, using good-quality input MSAs and phylogenetic trees is imperative for accurate ASR, irrespective of the method used. While MSA quality is still quite an elusive concept in general, here we assume that a good-quality MSA captures multiple-character (long) indels in a phylogenetically consistent way. Therefore, in our study, we use PRANK [Löytynoja and Goldman 2005], the phylogeny-aware tool which infers phylogenetically meaningful gaps by distinguishing insertions from deletions in a progressive manner on the tree.

Here, given accurate input data, we assess the systematic bias in PIP-based ASR by investigating the fragmenting of gaps in the inferred sequences at the ancestral nodes of the phylogeny. To test this, we present a large-scale analysis of protein orthologs from six mammalian species (human, three primates, and two rodents), taken from the popular orthologous protein database OMA [Altenhoff et al. 2021], as well as analysis of simulated data. We chose this specific phylogenetic dataset for two reasons. First, the mammalian species tree for these specific taxa is unambiguous and can be accepted as "true" (although the indel history is unknown, see [Nichols 2001]). Second, insertion and deletion biases in these species have long been a subject of interest, meaning that our findings can be interpreted in the context of current literature. For these data, we evaluated per-site insertion and deletion frequencies in different lineages and compared the gap distributions in the observed and inferred sequences.

To get a better understanding of ASR properties and potential biases under PIP, we proceed by analyzing simulated data. In our simulations, we mimic the OMA-based protein orthologous groups so that the results on real data can be compared to expected performance on very similar data where the truth is known. Our results suggest no significant difference in observed and inferred ancestral gap length distributions. This means that ARPIP tends to preserve the long indels from the input alignment in the inferred ancestral sequences. We also could confirm the well-documented deletion bias [Zhang and Gerstein 2003, Ogurtsov et al. 2004, Tao et al. 2007, Lin et al. 2017, He et al. 2019, Loewenthal et al. 2021].

4.3 Results

4.3.1 Results on mammalian data

We extracted and analyzed 12′088 orthologous protein groups, each containing one sequence from six *eutherian* mammals. Sequences in each orthologous group were aligned, and ancestral sequences were reconstructed given the inferred multiple sequence alignment (MSA) and the species tree (see data and methods; Fig. 4.12). For each site in an MSA, our ASR method ARPIP infers the most likely insertion and deletion history, allowing us to distinguish insertion and deletion events. Note that the reconstruction is done independently for each site, as in all other ASR methods. Therefore, we evaluated the number of inserted and deleted residues per site and per time interval rather than counting multiple residue events. This way of measuring indel rates is intuitively similar to substitution rates; therefore, it has a simple interpretation without having to account for the length of the full indel. Another advantage of this approach is that it makes it easy to evaluate the impact of indel events on sequence length over time. Note that here, we make a clear distinction between gaps and indels. Namely, gaps are the stretches of missing characters (gap characters "-") in a sequence resulting from the alignment. In contrast, indels are insertion and deletion events inferred via the ancestral reconstruction with ARPIP. Gaps in MSAs can appear due to several multiple-character insertions and deletions. Since ASR is performed independently at each site, gaps spanning multiple sites are described as a series of single-character indel events at several affected individual sites. To evaluate whether this assumption is reasonable during ASR, we study whether the ARPIP method preserves the distribution of gap lengths of the input MSA in the sequences reconstructed at ancestral nodes.

4.3.1.1 Comparing the number of inserted and deleted characters

1′267 of orthologous groups had no gaps in the inferred MSAs, presumably due to strong conservation. These groups were therefore excluded from the indel statistics presented here. For the remaining 10′769 orthologous groups, the total numbers of inserted and deleted residues on the species tree are visualized in Figure 4.1, and more detailed statistics are presented in Tab. 4.1. The *human* lineage had the lowest number of inserted and deleted characters, as well as overall gap characters in the sequences (4.5% of total sequence length). This is strongly contrasted by the *gorilla* lineage, which experienced the highest indel numbers among all studied species with 6.2% of its total sequences in MSAs consisting of gaps. Only the *chimp* and the *macaque* lineages had more gaps in their sequences, with 6.5% and 6.8%, respectively. These three primate lineages also had the longest on average gap lengths (on average 14.8 AAs for *chimp*, 15.2 for *gorilla* and 15.3 for *macaque*), compared to *human* (10.6) and all other lineages. *Hominini* ancestral lineage experienced the lowest number of inserted residues and indels over-

all, although this can be expected since this divergence corresponds to the shortest branch length on the species tree.

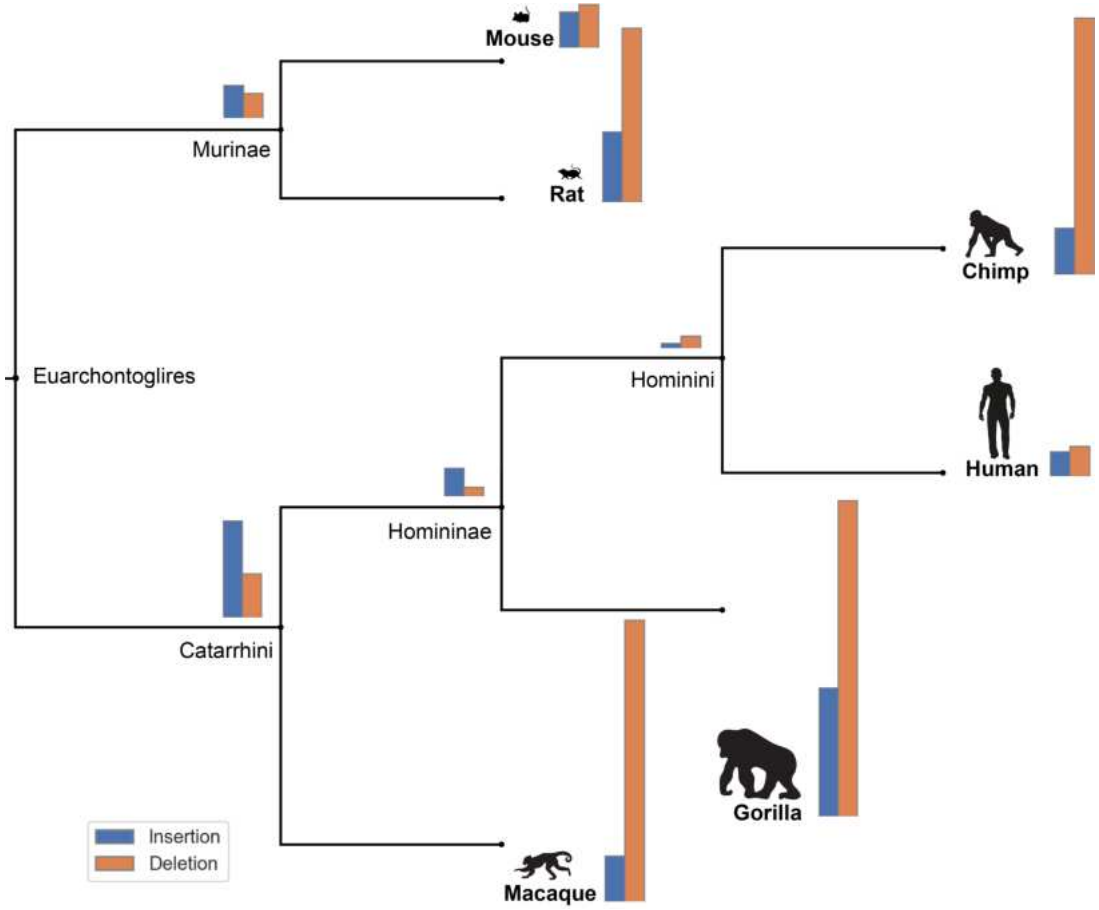


Figure 4.1: Number of indel events across studied species. *Gorilla* has the largest number of indel events per lineage while *Hominini* and *Homininae* have the lowest number of indel events, respectively (see Tab. 4.1).

Next, we calculated the insertion-deletion bias as the ratio between the numbers of insertion and deletion events (see Fig. 4.14 in Appendix). Overall, the number of deletions was larger than the number of insertions for all six extant lineages. The bias towards deletions was particularly strong in *macaque* (0.16) and *chimp* (0.18), but also well pronounced in *gorilla* and *rat* (both 0.41). In contrast, most ancestral lineages displayed a bias towards insertions, which was particularly pronounced in the *Homininae* (2.90) and *Catarrhini* ancestors (2.19).

4.3.1.2 Tracing the sequence lengths along the tree

Further, we investigated whether the observed deletion bias in extant lineages affects the sequence length dynamics across the species phylogeny. For each orthologous group, we computed Spearman correlation coefficients between sequence lengths (observed at the leaves or inferred at the ancestral nodes, gap characters removed) and the evolutionary distance from the

Lineage/Clade	Gap residues	Average gap length	Total gap	% gap residues	Average branch length	Ins	Del	Ins-Del bias
Human	316'690	10.6	29'998	4.5	0.004	16'907	20'465	0.82
Chimp	456'638	14.8	30'963	6.5	0.007	32'083	175'589	0.18
Hominini (Human, Chimp)	313'132	10.5	29'958	4.4	0.001	3'684	8'668	0.42
Gorilla	438'990	15.2	28'912	6.2	0.014	**90'145**	**220'987**	0.40
Homininae (Human, Chimp, Gorilla)	308'148	10.3	29'965	4.4	0.011	19'229	6'641	**2.90**
Macaque	**480'478**	**15.3**	31'365	**6.8**	0.017	31'389	191'131	0.16
Catarrhini (Human, Chimp, Gorilla, Macaque)	320'782	10.4	30'791	4.5	0.080	67'700	30'864	**2.19**
Mouse	357'149	9.7	36'777	5.0	0.030	25'067	30'165	0.83
Rat	423'379	11.3	37'640	6.0	0.034	48'601	119'929	0.40
Murinae (Mouse, Rat)	352'125	9.4	37'476	5.0	0.070	22'767	17'246	**1.32**

Table 4.1: Summary statistics of gaps and indels on mammalian data.

root to its corresponding node in the tree (evolutionary age). The majority of analyzed orthologous groups showed no significant correlations at a 5% significance threshold. Nevertheless, we observed significant correlations in 12.04% of orthologous groups with positive correlations for 459 genes and negative correlations for 838 genes (Fig. 4.2). This suggests that 7.78% of analyzed gene sequences had the tendency to shrink, while 4.26% had shown a tendency to grow.

4.3.1.3 Gap length distribution is preserved over time

We asked whether the gap distributions in the six observed sequences differed from those in the inferred ancestral sequences. The gap distribution in the inferred MSA of the six observed sequences results from the PRANK alignment and would, therefore, exhibit any inherent systematic biases of the PRANK method, if any. By analyzing whether a change in gap length distribution occurs at the inferred ancestral sequences, we aim to evaluate whether ARPIP tends to bias the distribution in a given alignment towards shorter gaps.

Such an effect is expected to be maximal at the root of the tree. Therefore, we compared the empirical distribution of gap lengths at the root with the distribution at the leaves over all OMA groups. Visually, the two distributions match (Fig. 4.3).

Furthermore, for each OMA group, we computed the mean gap lengths at the root and the mean gap lengths at the tips. The differences between the means are distributed around zero with a heavier tail in the positive range, which leads to an average difference of 4 characters, meaning

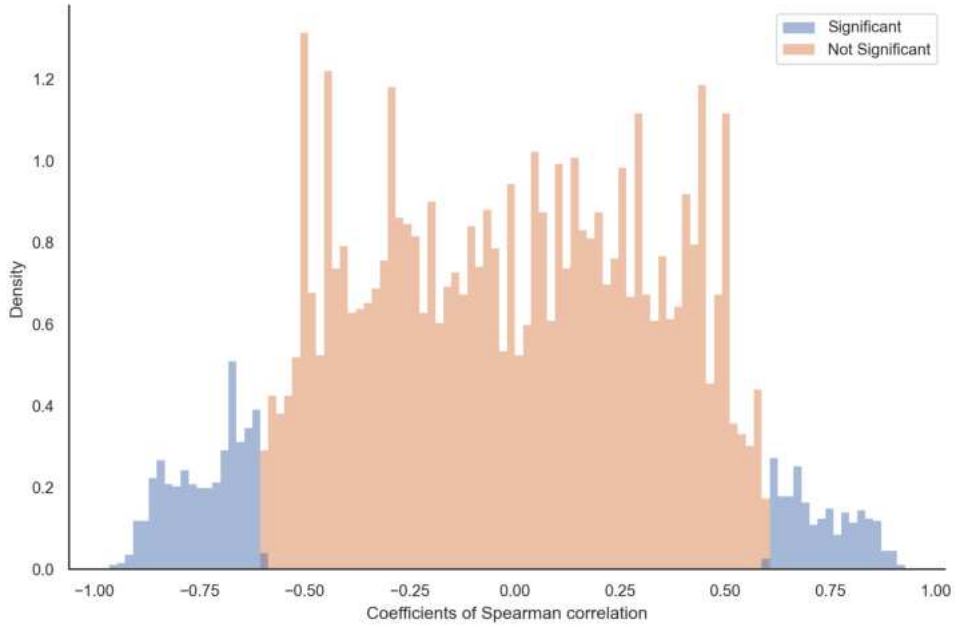


Figure 4.2: The distribution of Spearman correlation coefficients between sequence length (at the tips and root) and evolutionary distance from the root per OMA groups on six mammalian species.

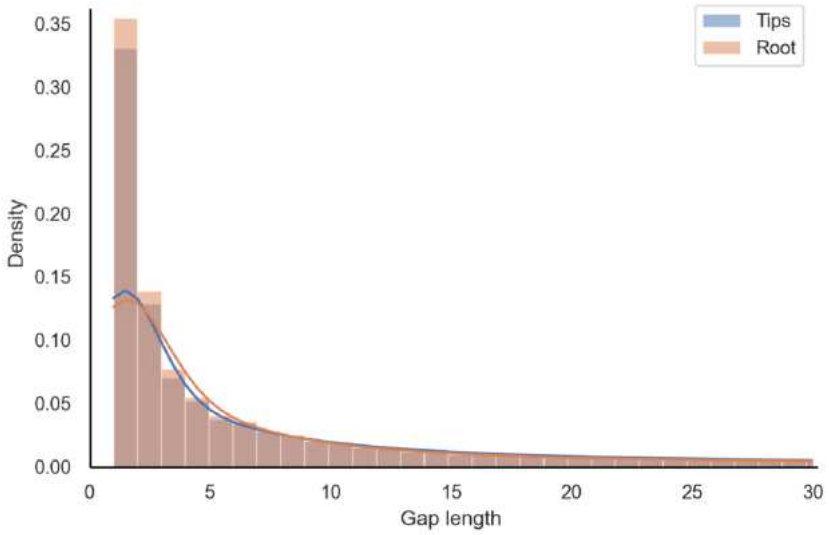


Figure 4.3: The empirical gap length distribution of tips vs. root on mammalian sequences. The plot is the histogram with 100 bins, while the upper bound of gaps is limited to 100 residues.

that gaps at the tips tend to be around 4 characters longer (Fig. 4.4).

4.3.1.4 Inserted segments are longer than deleted segments

Finally, we compared the empirical distributions of multiple-character insertion and deletion events over time on the phylogeny. Figure 4.5 depicts that the empirical distributions of inser-

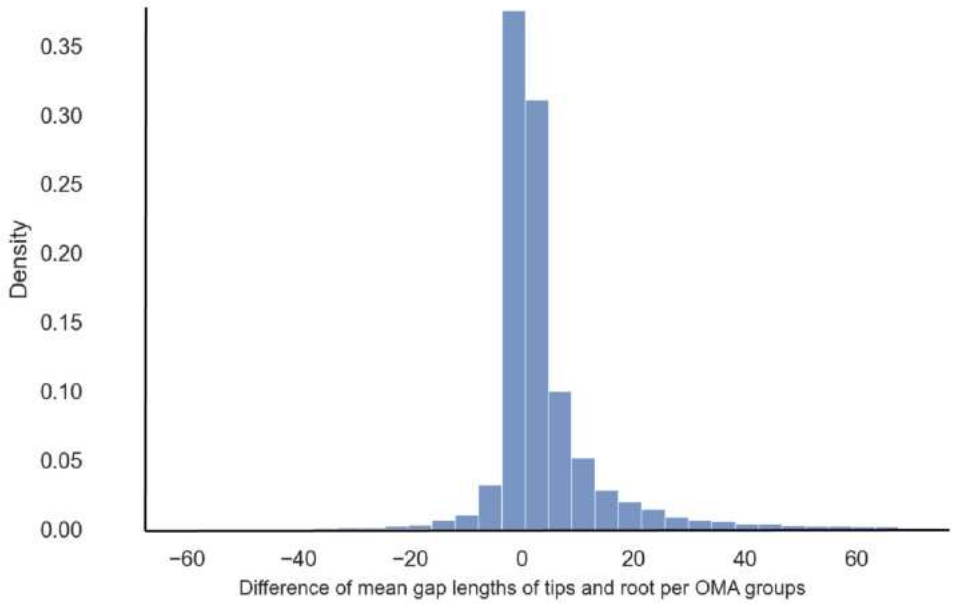


Figure 4.4: Paired difference of mean gap lengths per OMA groups on mammalian data (with 100 bins).

tions and deletions are consistent with the empirical gap length distribution as single-character events are the most frequent, and their frequency decreases as the length of the event increases. In addition, we observed that insertions tend to be significantly longer than deletions, where the mean insertion length was 17.81, while it was 8.67 for deletion events.

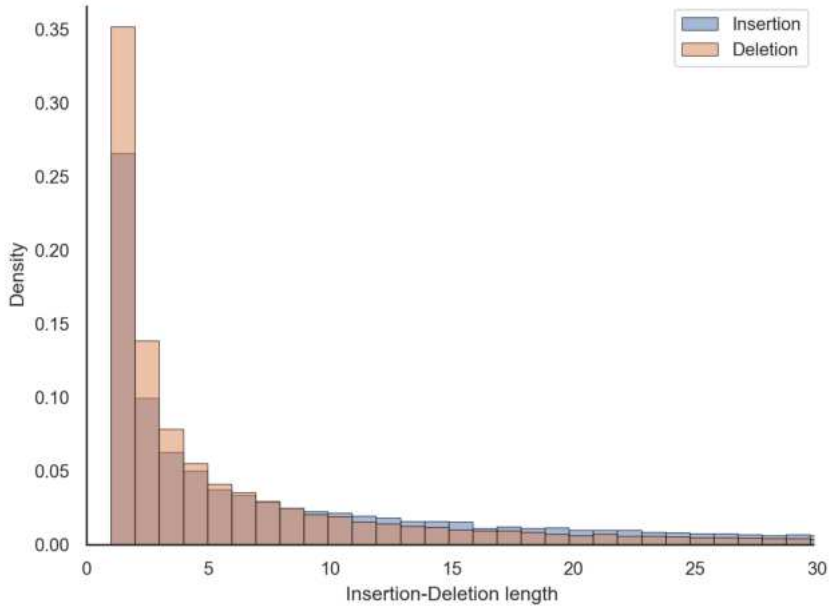


Figure 4.5: The empirical distribution of inserted vs. deleted segment lengths. The plot is the histogram with 100 bins, while the upper bound of gaps is limited to 100 residues.

4.3.2 Results on simulated data

To study ARPIP under fully controlled conditions, we have simulated sequences with INDELible. To set realistic parameters, we sampled 1000 random OMA groups. For each sampled OMA group, we used the corresponding PhyML tree to evolve a replicate on it, with the root sequence length of 1000 AAs, indel rate of 0.1, and indel lengths distributed according to the Zipfian distribution with exponent 1.7. INDELible's maximum indel length parameter was set to the length of the longest gap in the PRANK MSA of the OMA group in question. We supplied the true simulated MSA of the observed sequences to ARPIP for all the analyses.

4.3.2.1 Reconstruction accuracy

On simulated data, ARPIP inferred a positive insertion-deletion bias in all nodes of the trees; i.e., more individual characters were inserted than deleted (Appendix Fig. 4.15). It correctly reconstructed more than 98% of ancestral residues, resulting in 90% correctly inferred ancestral columns (Appendix Tab. 3.1). The average precision[1] in gap character inference was 94%, with a recall[2] of 97%. We divided the simulation results according to the F-score (a measure of predictive performance defined as the harmonic mean of precision and recall) in gap retrieval into "optimal" (132 samples with F-score $\geq 99\%$) and "sub-optimal" (858 samples with F-score $< 99\%$). Figure 4.7 shows the branch length distributions for the two classes. Note that for the sub-optimal samples, this distribution is near zero. For these samples, we observed a lower accuracy in gap reconstruction. Indeed, shorter branches provide less information, and we expect larger variances and lower accuracy. Likewise, this happens also for extremely high divergences when the signal becomes saturated. Furthermore, the insertion probability in PIP is proportional to branch lengths. Thus, the choice of insertion points also depends on the relative branch lengths of the phylogeny. Figure 4.6 shows the ROC curve points for each sample (and not just one point, i.e. the average).

4.3.2.2 Tracing the sequence lengths along the tree

Analogous to the real data analysis above, we correlated the sequence length without gaps in each node with the node's evolutionary age for each replicate. Again, the majority of the Spearman coefficients were not significant at the threshold of 5%. Among the 14.95% significant ones, we observed 93 positive and 55 negative correlations (Fig. 4.8). Contrary to the real data, here, the majority of the significant replicates tended to grow, while 37% were shrinking. This is consistent with the positive indel bias.

[1]The percentage of correctly inferred gap characters among all inferred gap characters
[2]The percentage of correctly inferred gap characters among all true gap characters

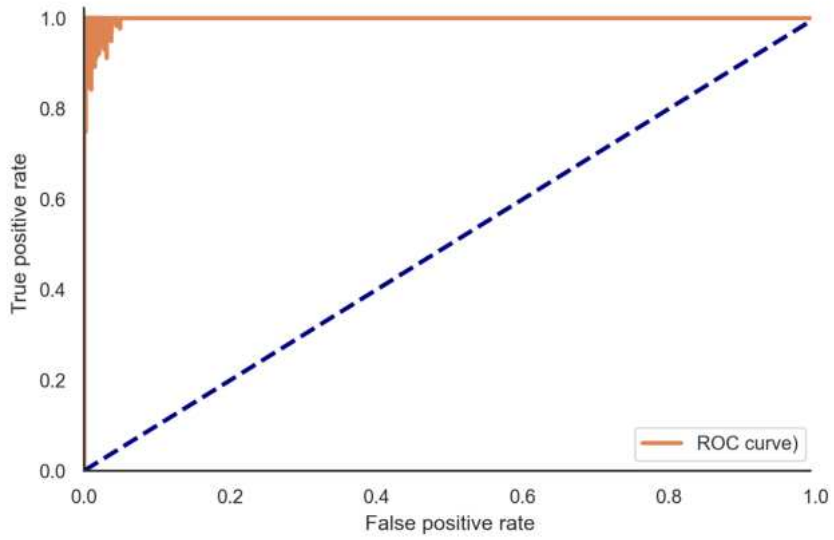


Figure 4.6: ROC curve: true positive (recall or sensitivity) vs. false positive (1-specificity) rates at the ARPIP gap estimation.

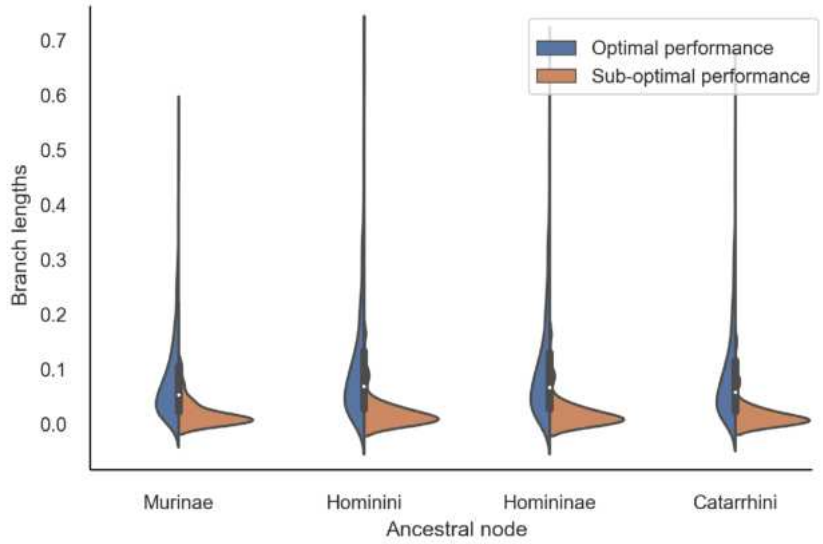


Figure 4.7: Distribution of ancestral node branch lengths in the simulated data, grouped by inference performance.

4.3.2.3 Gap length distribution is preserved over time

Next, we asked if the gap length distribution in the inferred ancestral sequences differed from the true distribution, i.e. the one generated by the simulation. The two distributions match (Fig. 4.9) and have a Kullback-Leibler divergence of 4.42×10^{-5}. According to the PIP model, we expect sequence lengths to be preserved, meaning not shrinking nor growing. Furthermore, there seems to be no decline of gap lengths towards the root of the tree, as the gap length distribution inferred at the root of the tree matches the distribution in the observed sequences at

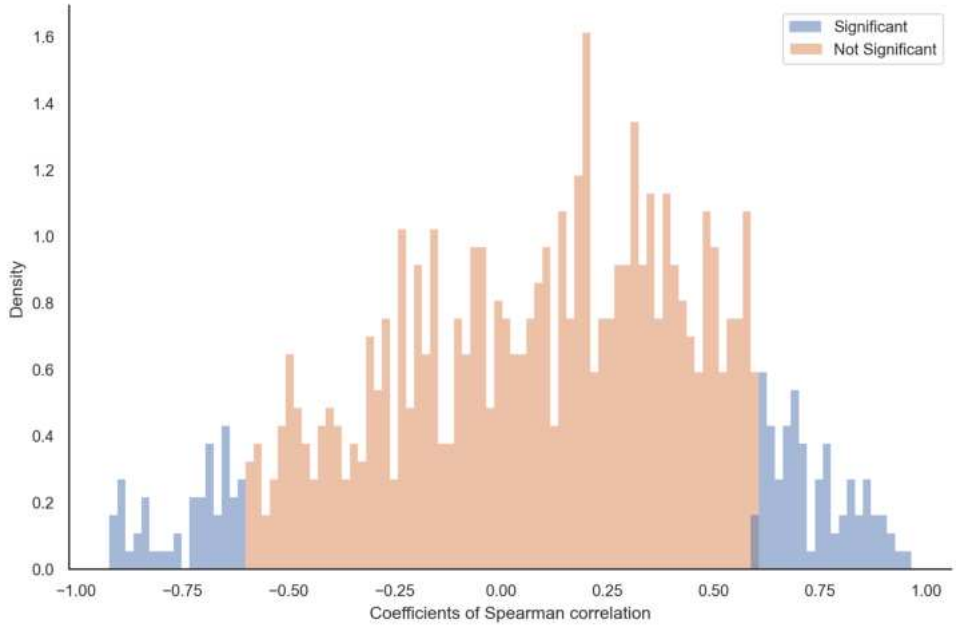


Figure 4.8: The distribution of Spearman correlation coefficients between sequence length (at the tips and root) and evolutionary age per OMA group on simulated data.

the leaves (Fig. 4.10). Note that in contrast to the real data case above, where the gaps at the leaves were inferred by PRANK, here we were able to compare to the true (simulated) MSA.

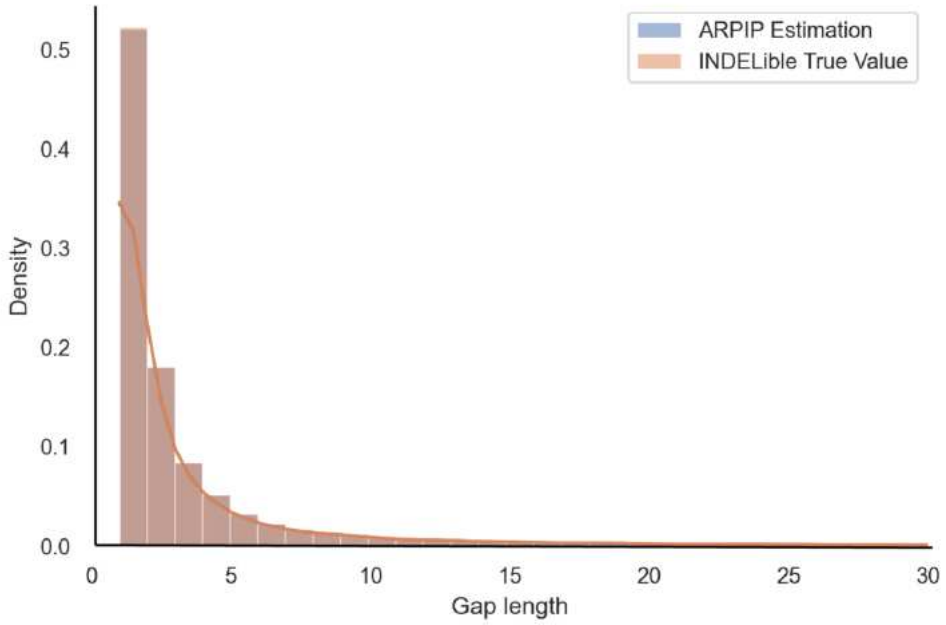


Figure 4.9: Overlapped distributions of gap lengths from ARPIP inference and INDELible true values.

To further quantify the difference between simulated and inferred distributions, we computed the mean gap lengths at the root and the mean gap lengths at the tips for each of the 1000

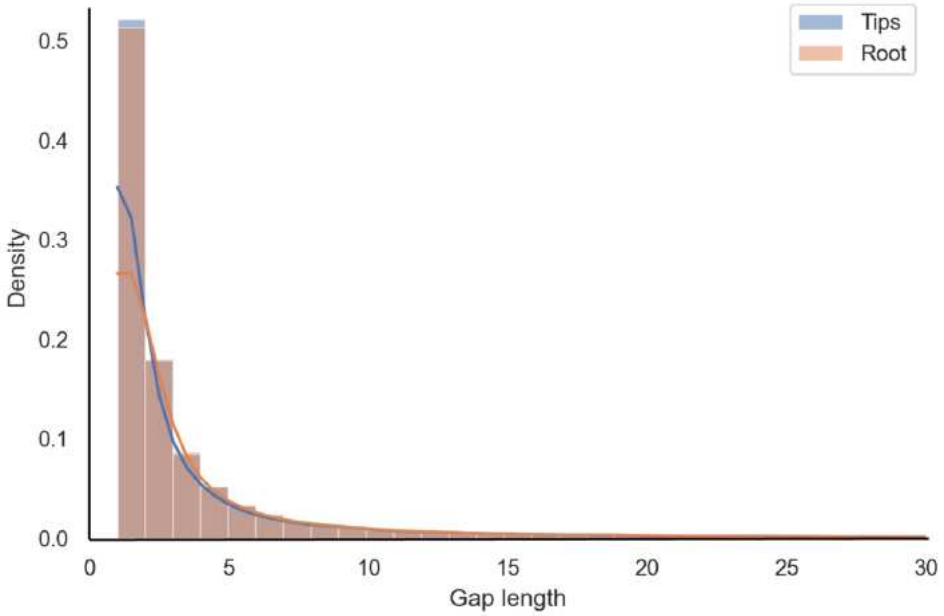


Figure 4.10: Empirical gap length distribution at the tips vs. the root in simulated sequences as a histogram with 100 bins with a cut-off at 100.

replicates. Analogously to the mammalian data, the differences between the means were symmetrically distributed around zero (Fig. 4.11). The differences were not statistically different from zero (Mann-Whitney test, $p = 0.67$; two-sample t-test, $p = 0.997$).

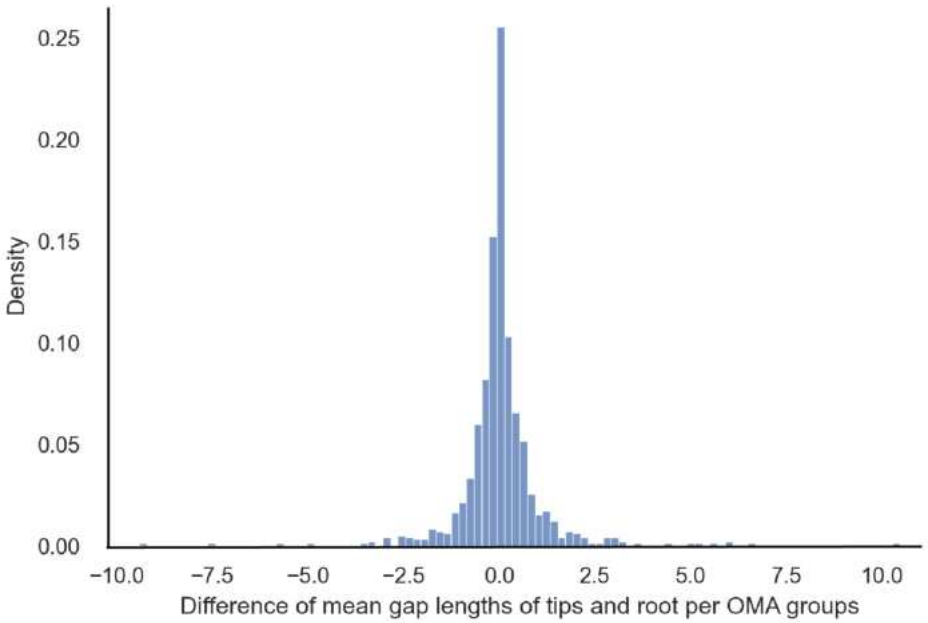


Figure 4.11: Paired difference of mean gap lengths per OMA groups on mammalian data (with 100 bins).

In summary, our simulation findings corroborate the results from real data. ARPIP preserves the gap lengths from the input alignment.

4.4 Discussion and conclusions

Until recently, state-of-the-art ASR methods focused on inferring ancestral characters. Indels were often mishandled – either by removing gappy MSA columns, treating gaps as ambiguous characters [Yang 2007], or reconstructing ancestral gaps with ad-hoc indel methods like "indel coding" [Ashkenazy et al. 2012]. Further, such methods typically do not easily distinguish between insertions and deletions. Unlike previous approaches, ARPIP reconstructs insertions and deletions independently and uses the evolutionary indel model PIP. However, PIP only describes single-character indels.

In contrast to ASR, methods for MSA inference are more advanced with respect to allowing for long indels. One of the most advanced aligners is PRANK; it uses the phylogeny to distinguish insertions from deletions and, thus, infers phylogenetically meaningful long indels. All current ASR methods take an MSA as input. Here, we have shown on real data (with PRANK alignments) and by simulation (with the true simulated MSAs from INDELible) that the ancestral estimates by ARPIP preserve the long indel structure present in the MSA. This surprising result can partly be explained by the fact that under PIP the insertion and deletion points of a site only depend on the gap patterns (i.e. the presence and absence of gaps), and are independent of the character states [Jowkar et al. 2023]. Neighboring sites with identical gap patterns form long indels and lead to identical indel histories (see, for example, Appendix 4.8.2). Further studies will be needed to quantify how differences in neighboring gap patterns affect long indel preservation. Based on ARPIP's strong performance, we hypothesize that minor pattern differences will still preserve most long indels. Also, analogously to other phylogeny-related inference problems [Bergsten 2005] our simulations showed that short branches lead to a lower accuracy in gap reconstruction.

Furthermore, in line with the biology [de Jong and Rydén 1981] and previous bioinformatics studies [Kuo and Ochman 2009, Zhang and Gerstein 2003, Loewenthal et al. 2021], we found that deletions are more frequent than insertions. Such deletion bias has been detected across the whole tree of life and has multiple possible evolutionary explanations. For example, [He et al. 2019] suggests that even strictly balanced insertion and deletion rates result in a linearly increasing genome size through time rather than a completely fixed genome size. The authors attribute this effect to the fundamental asymmetry of indels, as insertions produce more characters available for deletion, while deletions reduce the total number of characters, resulting in fewer deletable ones. The authors suggest that while the huge variety in genome sizes among

species seems to require exponential size growth, the effective insertion bias cannot act for prolonged periods of evolutionary time. Consequently, the mechanisms producing larger genome sizes only act sporadically and are likely to be removed in the long term, making them very difficult to detect by looking into existing genomes. On the other hand, the commonly detected deletion bias could be an effect similar to the pull-of-the-present effect in phylodynamics, where younger lineages show seemingly higher birth/lower death rates [Nee et al. 1994]. This effect stems from the fact that we are observing a snapshot of the evolutionary history that is cut off from the future, meaning that while some of the present-day lineages might go extinct, they have had less time to do so than older lineages and thus are more likely to have gotten sampled. In essence, this would mean that deletions might appear more frequently in the present sequences, but only because they have not yet been fixed in the underlying population at the moment of sampling.

Finally, until now, studies on indel length distributions have lumped the insertions and deletions together, often just inferring gap length distributions. As a step forward, we suggest inferring separate distributions for insertion and deletion lengths. Our findings from mammalian data strongly point to longer insertion lengths than deletion lengths. Further, given the higher prevalence of deletions and the remarkable uniformity of protein length distribution across the tree of life [Nevers et al. 2023], it is conceivable that the two distributions differ, with deletions lengths having a smaller mode than insertions. Recent work from Tal Pupko's lab is a notable step in the direction of inferring indel length distributions based on event reconstruction [Wygoda et al. 2024].

4.5 Data and methods

4.5.1 Sequence acquisition and alignment

First, we used the OMA database [Altenhoff et al. 2021] to obtain orthologous protein sequences so that each orthologous group (OMA group) contained one sequence from each of six mammalian species, namely *human*, *chimp*, *gorilla*, *macaque*, *mouse*, and *rat*. The OMA database is known for its higher precision but lower recall compared with the majority of other methods [Altenhoff et al. 2019b; 2021]. A corresponding species tree was extracted from the Ensembl Compara v. 105 [Yates et al. 2020] by pruning a larger mammalian tree to the six species considered in this study (see Fig. 4.13). This species tree was then provided as a guide tree for reconstructing multiple sequence alignments (MSAs) using PRANK+F, a phylogeny-aware progressive aligner distinguishing insertions from deletions [Löytynoja and Goldman 2005]. For each reconstructed MSA, branch lengths on the species tree were re-optimized by maximum likelihood with PhyML v3.3.20211231 [Guindon et al. 2010]. Codeml from PAML

[Yang 2007] was used in a few cases where PhyML optimization has failed. Finally, a refined PRANK MSA was inferred for each orthologous group using a species tree with re-optimized branch lengths as a guide tree. The WAG AA substitution model [Whelan and Goldman 2001] was used in all analysis steps, including the ancestral sequence reconstruction described below.

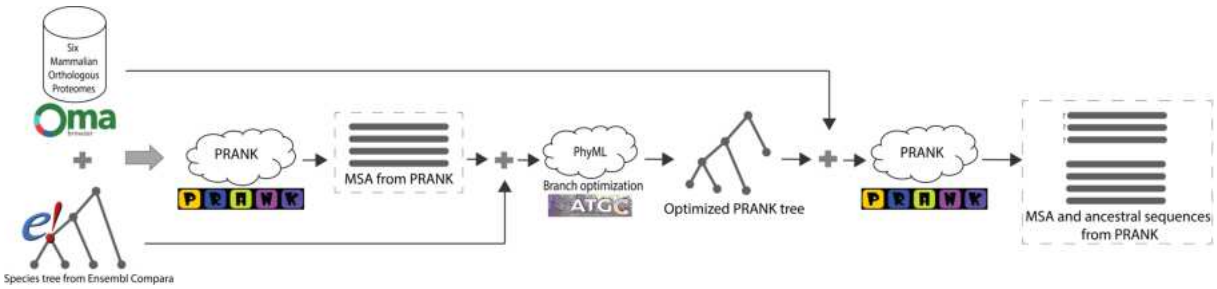


Figure 4.12: Algorithmic data acquisition pipeline.

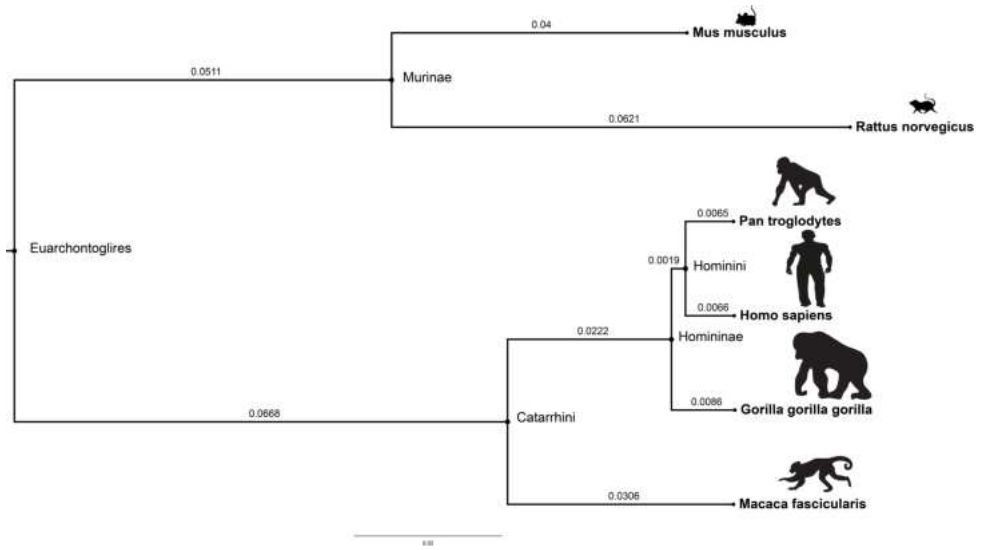


Figure 4.13: Illustration of the guide tree extracted from 43 *eutherian* mammals. The branch lengths were estimated using pairwise MSA in Ensembl Compara v.105.

4.5.2 Ancestral sequence reconstruction

The refined MSA inferred above was used to infer ancestral sequences at all nodes of the species tree with optimized branch lengths. For this purpose, we used our recent method implemented in ARPIP [Jowkar et al. 2023]. Evolutionary changes on a phylogeny are described via the PIP model [Bouchard-Côté and Jordan 2013], where insertions follow a Poisson process, while substitutions and deletions follow a continuous-time Markov model with an absorbing state. The ARPIP method includes two main steps. First, the method infers the most probable indel scenario on a given phylogeny, independently for each column of a given MSA. Next, similar to FastML [Pupko et al. 2000], ancestral characters are reconstructed on a subtree of the given

phylogeny obtained by pruning it to the inferred indel scenario. For ASR analyses, the root was placed on the internal branch connecting the *primate* and *rodent* clades. Then, midpoint rooting was used to define the location of the root on this branch.

4.5.3 Simulating data

We simulated 1000 data sets with INDELible [Fletcher and Yang 2009]. To set realistic parameters, we sampled uniformly at random 1000 OMA groups and extracted the corresponding PRANK MSAs and species trees with PhyML-optimized branch lengths (as described above). For each sample, we simulated a replicate on the PhyML tree using a sequence of 1000 AAs at the root. We use a Zipfian indel length distribution with $\alpha = 1.7$, a maximum indel length equal to the maximum gap length of the OMA group in question, and an indel rate of 0.1. Sequence lengths in the simulated samples ranged between 336 and 1730 AAs, while the gap lengths ranged from 1 to 1451 characters. Around 1% of simulations produced biologically unrealistic sequences with extremely long gaps, for example, the sample with a 1451 character long gap. Such samples would be considered noisy in real datasets (possibly due to sequencing errors) and were thus also removed from the simulation analysis before evaluating reconstruction performance. Only four simulated samples contained no gaps at all and were also removed from analysis. The final simulated dataset contained 786 to 1371 AA long sequences and the gap lengths ranged from 1 to 235 characters.

We provided the true MSA from the simulation and the PhyML tree (i.e. true tree) to ARPIP for ancestral reconstruction.

4.6 Funding

This work was funded by the Swiss National Science Foundation (SNSF) grant no. 31003*A*_176316 and no. 315230_215379 to M.A. The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data, nor did it in writing the manuscript.

4.7 Available data and scripts to study indel pattern

This manuscript is accompanied by the scripts used to produce the results. The experimental data used in this manuscript is freely available from `https://doi.org/10.5281/zenodo.10798097`. The Python scripts used for data processing and analysis are also available at `https://github.com/acg-team/single-char-indel-ASR-preserves-long-indels`.

4.8 Appendices

4.8.1 Appendix 1: Tables and figures

4.8.1.1 Tables related to the accuracy of reconstruction on the simulated data

We report the average accuracy over all the samples.

Metric	Consistency (%)
Proportion of correctly inferred ancestral characters	97.88 ± 2.01
Proportion of correctly inferred ancestral columns	90.35 ± 2.01
Proportion of correctly inferred ancestral amino acids (i.e., excluding gaps)	97.75 ± 2.55
Gap precision	94.27 ± 5.37
Gap recall (sensitivity)	96.99 ± 3.97
Gap F-score	95.46 ± 3.37
Gap specificity	99.29 ± 1.17

Table 4.2: ARPIP performance in simulation. All metrics include the root sequences. They have been computed for each sample individually. We report the averages over the samples.

Lineage/Clade	Gap consistency/accuracy (%)		
	Precision	Recall	F-score
Murinae	98.74 ± 5.24	99.93 ± 1.10	99.31 ± 1.60
Hominini	99.99 ± 0.11	$99.9995 \pm 0.02 (100\%)$	$99.9970 \pm 0.06 (100\%)$
Homininae	99.98 ± 0.20	99.94 ± 1.51	99.95 ± 0.97
Catarrhini	98.48 ± 1.75	99.96 ± 0.60	99.71 ± 0.99
Euarchontoglires	77.71 ± 17.22	85.49 ± 17.46	79.44 ± 15.07

Table 4.3: ARPIP performance in gap character inference by simulation. Performance is shown individually for each internal node.

4.8.1.2 Indel bias plots for the mammalian and simulated data

Figure 4.14 and Figure 4.15 representing indel bias of mammalian and simulated data.

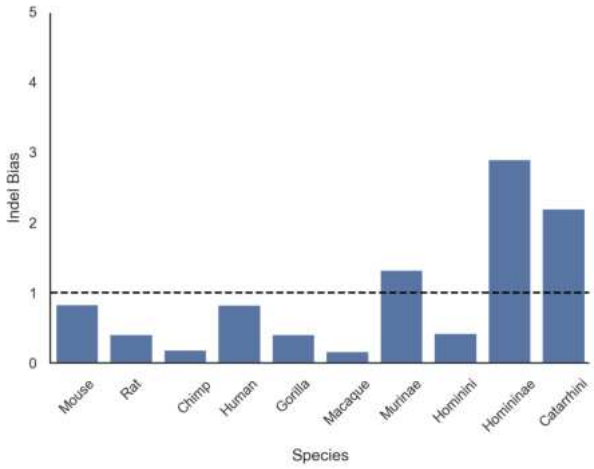


Figure 4.14: Indel bias (ratio of insertion to deletion events) in mammalian data. A ratio of less than one indicates a bias toward deletions.

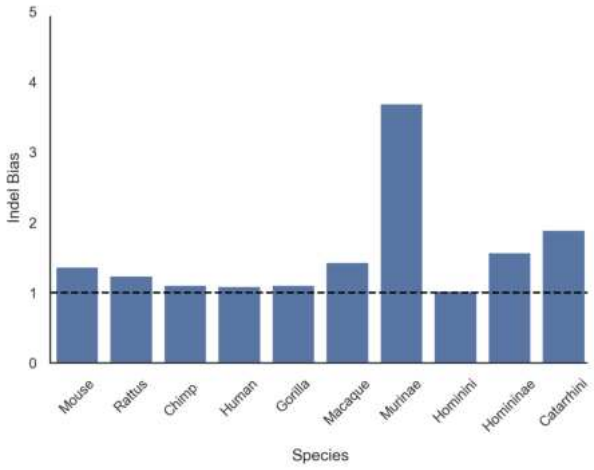


Figure 4.15: Indel bias (ratio of insertion to deletion events) in simulated data. A ratio of less than one indicates a bias toward deletions.

4.8.2 Appendix 2: Study of example reconstructions on simulated data

To get a better intuition for the performance of indel reconstruction under PIP, we have selected two samples from the pool of simulated data for closer examination. The sample s_1 is among the data with the lowest gap retrieval performance, while the second sample s_2 is a sample with a relatively good gap retrieval score.

4.8.2.1 Sample 1: Sub-optimal performance

We have selected two samples from the pool of simulated data to study the performance of gap reconstruction of ARPIP. Sample s_1 is among the samples with the lowest F-score. For s_1, the F-score is 72.86%, while precision and recall are 100% and 57.31%, respectively. This means that all the inferred gaps were correct, but only around half of the gap characters were inferred. The inference accuracy at the root was the lowest not only in this sample but also in all the samples from the simulated dataset (see Tab. 4.3). Figure 4.17 visualizes a segment of s_1 to investigate ASR performance and gap patterns.

Figure 4.17 highlights the inferred and true ancestral sequences for four regions of interest. Region R1 depicts five independent insertion and deletion events. Each insertion happened at the root, followed by deletion at the *macaque* taxon. Region R1 does not affect the ancestral gap length distribution, but this typical case happens for a single stretch of gaps at the taxa node. Similarly, region R2 occurs when a single residue in an MSA column exists. A single insertion at the taxa node usually represents a single residue insertion event. This inserted site will show up as a gap in all ancestral nodes, affecting the gap length distribution at the ancestral node, while in reality this site never existed in the ancestor.

Region R3 contains multiple long gaps within both ancestral and taxa species. In this case, as the neighboring site across both ancestral and descendent nodes has the same gap pattern, ARPIP infers the same indel scenario given fixed model parameters. A single insertion at the tree's root is followed by deletion at the *Murinae* branch.

In s_1, the gap reconstruction accuracy in the root node is very low due to low recall, meaning that ARPIP reconstructs a small fraction of the gaps in the root node. The cause for low gap character retrieval rates remains to be explained. Figure 4.16 shows different indel scenarios for a constant MSA column with respect to the branch length of the tree.

Region R4 is a masking indel event of region R3, as we have an insertion event at the branch leading to the node *Murinae*. This is a single-site indel event affecting the ancestral and descendant gap distribution. Notice that we have a single insertion at node *Murinae* without any deletion events.

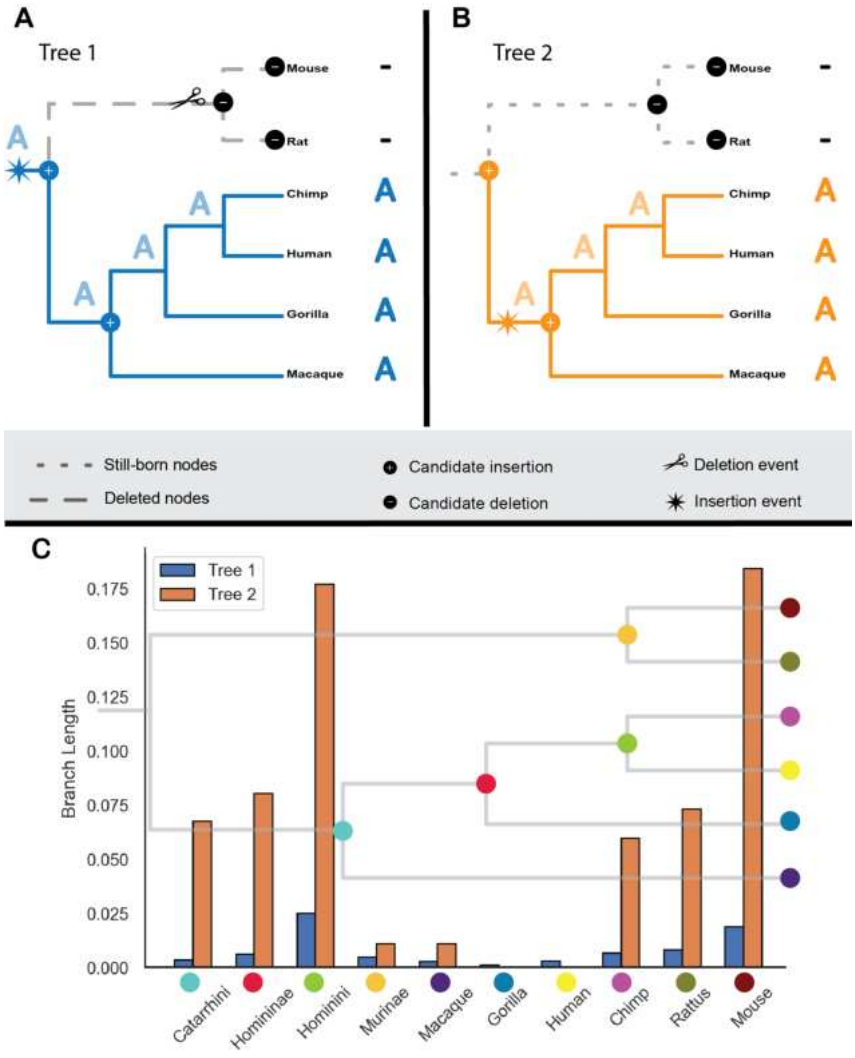


Figure 4.16: A, B) Two different indel scenarios for a single MSA with various branch lengths. C) Histogram of branch lengths of two selected simulated samples.

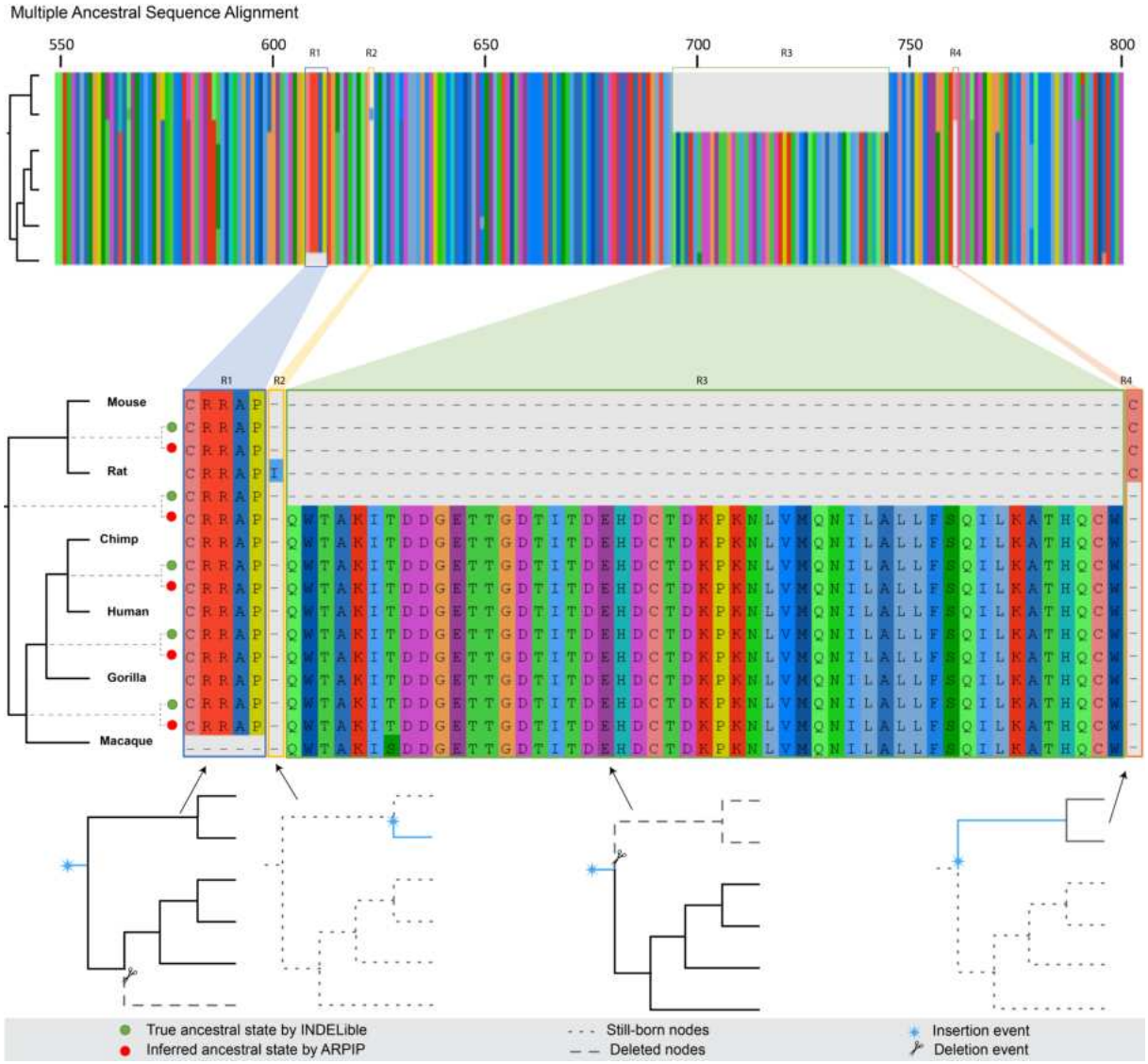


Figure 4.17: Multiple ancestral sequence alignment of ARPIP inference and INDELible true ancestral states for sites $550-800$ of sample s_1. The indel inference for each site is shown at the bottom of the figure.

4.8.2.2 Sample 2: Optimal performance

In addition, we have selected sample s_2 with an overall F-score of 91.36%, resulting from 84.10% precision and 100% recall. This implies that all the gaps were inferred correctly, while a fraction of non-gap characters were falsely inferred as a gap. Figure 4.18 illustrates that ARPIP performs well in inferring ancestral sequences despite the complex gap pattern, with 93.42% overall reconstruction accuracy. Moreover, sample s_2 performs relatively well at the root node compared to sample s_1. Figure 4.18 shows the gap pattern in two selected neighboring regions (R1-3) and (R4-6).

The PIP model tends to place the insertion events at the root because the Poisson process ini-

tiates at the tree's root. Regions R1 and R3 have a repeated insertion at the root followed by a single deletion event at the *rat* taxon. A neighboring region denoted by R2 has an additional gap between the regions mentioned. ARPIP can adapt the indel event for this specific site while preserving the gap distribution for the other two regions. The gap pattern in these three regions did not affect the gap distribution of ancestral nodes.

The transition from gap pattern R1 to R2 (sites $606-607$) and R2 to R3 (sites $607-608$) suggests that introducing a new gap in another node would have minimal impact on gap inference. The transition from R4 to R5 (sites $656-657$) or R5 to R6 (sites $668-669$) shows the ARPIP can preserve gap distribution for the long ancestral gaps. The results suggest that ARPIP performs exceptionally on neighboring sites with long gaps but suboptimally at the root.

Neighboring segments R4-R6 show two different indel event patterns. We infer that the R4 and R6 segments have an insertion at the *Catarrhini* node, and the R5 segment has an insertion at *Homininae*, without any deletion events at these sites. These three neighboring regions would affect both the ancestral and descendant gap patterns. Like in sample s_1, region R5 separates R4 and R6 without negatively affecting the gap inference. This example shows that ARPIP is relatively good at preserving gap patterns in the neighboring sites.

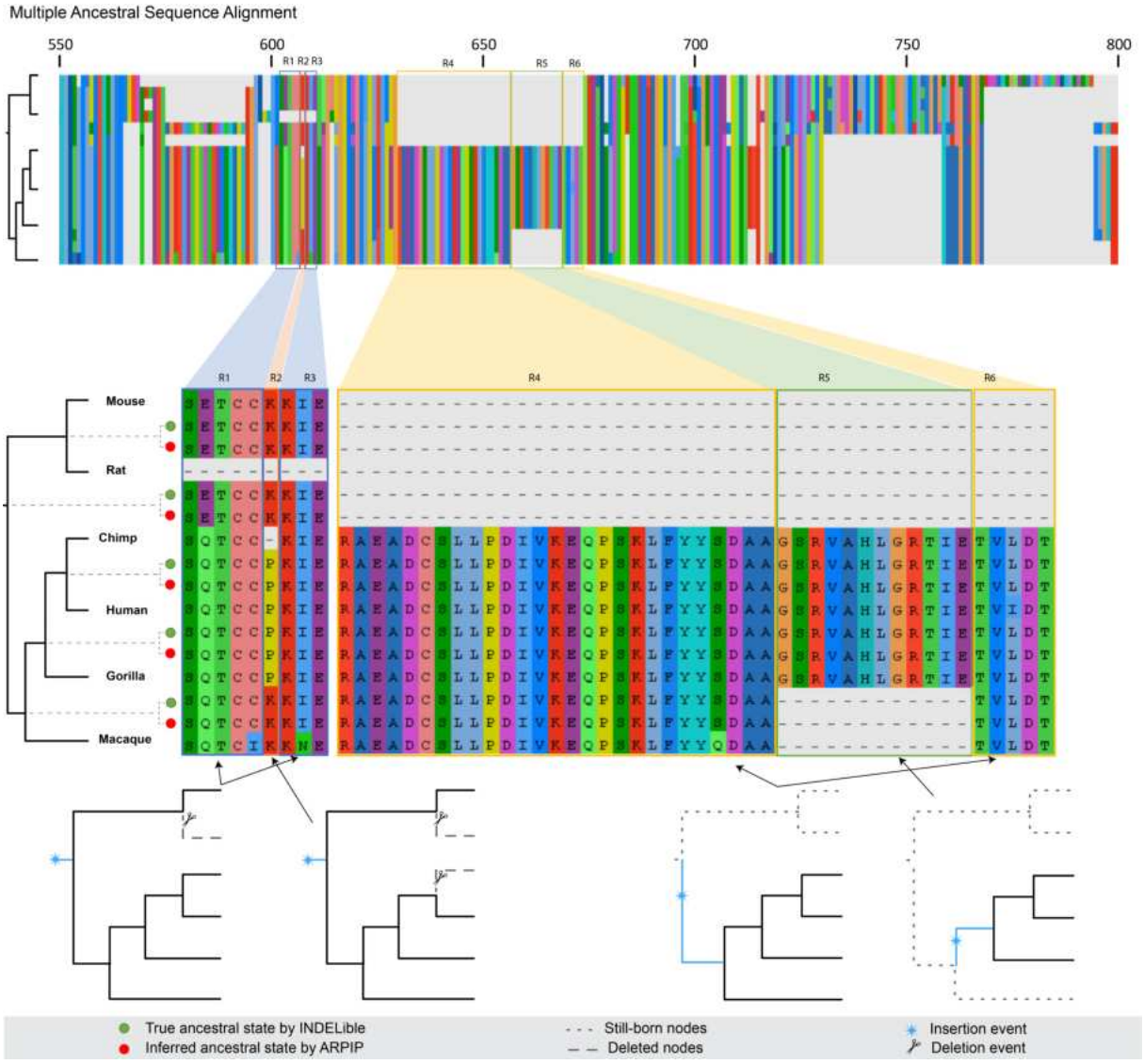


Figure 4.18: Multiple ancestral sequence alignment of ARPIP inference and INDELible true ancestral states for sites 550 – 800 of sample s_2. The indel inference for each site is shown at the bottom of the figure.

Chapter 5

A Probabilistic Solution to Measure the Uncertainty of the Ancestral Sequences for Squamates Neural Retina Leucine Zipper Transcription Factor

This chapter is a technical report (unpublished). The primary objective of this work was to enhance the handling of uncertainty in ASR. As an application, the neural retina leucine zipper transcription factor was selected due to its interesting application in the visual system and based on our current collaboration. Please notice that the material used in this report is confidential until its publication. The conceived and designed experiments, including the data and the case study scenario questions, were provided by Prof. B.W. Chang and E. Dong. The computational experiments and their technical report were done by the present researcher. Currently, our collaborators are writing a manuscript (not presented for this thesis), which will include the analysis of data discussed in this chapter. I will be a co-author of this manuscript.

A Probabilistic Solution to Measure the Uncertainty of the Ancestral Sequences for Squamates Neural Retina Leucine Zipper Transcription Factor

Gholamhossein Jowkar[1,2,3,*], Manuel Gil [2,3], and Maria Anisimova [2,3]

[1] *University of Neuchâtel, Institute of Biology, CH-2000 Neuchâtel, Switzerland*
[2] *Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*
[3] *Zurich University of Applied Sciences, LSFM, ICLS, CH-8820, Wädenswil, Switzerland*

**Gholamhossein Jowkar, ZHAW, LSFM, ICLS, Schloss 1, 8820 Wädenswil, Switzerland, E-mail: jowk@zhaw.ch*

Abstract

Statistical ancestral sequence reconstruction (ASR) models aim to provide an accurate most probable ancestor (MPA); in practice, these methods frequently yield results associated with sizable uncertainty. This is problematic in vitro due to the significant expense of synthesizing and expressing such sequences. One possible resolution for this matter is that the biological expert can decide on the reconstruction ambiguity using the probability profile of each reconstructed ancestral state. This would allow one to consider the variability of ancestral state inference rather than the single MPA estimate. To resolve this issue, we have proposed two probabilistic approaches, normalized and naive posterior probability profiles under the Poisson Indel Process (PIP) model and using the empirical Bayes ASR.

This is illustrated on a dataset consisting of sequences from neural retina leucine zipper (NLR) transcription factors (TFs) from 81 species. This protein is involved in developing rod photoreceptors in vertebrates. The sequences were provided by Prof. B. W. Chang's team based at the University of Toronto. With the help of the two variants to compute ASR probability profiles, we will support the experimental reconstruction of squamate ancestor proteins in the laboratory. The experiments will be conducted by the team of Prof. Chang. This work will also help to troubleshoot our approach and learn about potential caveats.

In this project, the phylogenetic information was extracted using state-of-the-art methods from the selected NLR TFs. Using our recently developed ASR method ARPIP, we have inferred the ancestral sequence and their probability profiles for selected vertebrates' ancestral nodes.

Keywords: joint ancestral sequence reconstruction, neural retina leucine zipper transcription factor, squamates ancestral sequences, Poisson indel process, uncertainty of inferred ancestral sequences.

5.1 Introduction

Molecular evolutionary knowledge about the retina within the eye can help us to find the possible cause of visual impairment and its cure. The retina, within the eye, is a highly evolved organ that captures and processes this visual information. Among the cells carrying information from the eye to the brain, photoreceptor cells initiate the vision process. Photoreceptors are highly specialized primary sensory neurons that sense light, while their loss or dysfunction results in visual impairment [Wang and Cepko 2016]. The retina consists of two types of photoreceptor cells: rods and cones. Rods are sensitive to light levels and help us get good vision in low light, while cones detect different colors and require brighter environments. Neural retina leucine zipper (NRL) is an essential transcription factor (TR) in the cellular and functional maintenance of rod photoreceptors in the mammalian retina [Hao et al. 2011]. In many human retinal diseases, visual impairment is due to photoreceptor dysfunction. The current study could provide insights for designing new treatment strategies, such as photoreceptor replacement therapy for human retina-related diseases.

NRL TF is viewed largely as a critical determinant of rod photoreceptor cell fate [Montana et al. 2011], while misexpression in cones might transform them into rods [Wang and Cepko 2016, Swaroop et al. 2010, Schott et al. 2016]. There are some theories about the evolution of rods in vertebrates [Simoes et al. 2016] and evidence that rod cells can contribute to the color vision in snakes [McKee et al. 1977]. The transmutation theory suggests that the photoreceptors in the group of snakes and lizards could have transformed from one type to another (i.e., cones to rods) a long time ago [Walls 1934]. By generalizing the transmutation hypothesis, squamates NRL TF are ideal groups to study photoreceptor fate during the development of vertebrates due to their highly variable photoreceptor morphologies.

There are insufficient studies on the evolutionary forces affecting NRL TF, including indels. On the one hand, existing studies on the evolution of the vertebrate visual system mostly focused on amino acid (AA) substitution mutation (for example, evaluating positive selection) while removing alignment gaps representing indels or treating them as missing values [Simões et al. 2016]. On the other hand, small-scale mutations like indels have an important role in gene expression patterns within mammals, as TFs are known to be highly enriched with indels [Ribeiro-dos Santos et al. 2015]. In this research, we will study these driving evolutionary forces by investigating the small-scale mutations (substitutions and indels) in NRL responsible for the transmutation of photoreceptors in the studied group.

It is also known that the inference of ancestral sequences involves a level of uncertainty, while maximum likelihood (ML) based methods overcome this issue to a certain degree. From an

inferential point of view, inferring an accurate MPA is the objective; however, it has been known that the inferred ancestral sequences are biased by errors in the alignment and tree inference [Pollock and Chang 2007]. The biological expert can resolve the uncertainties in reconstruction using probability profiles. For example, an expert can choose one AA over the other based on some aspect of the gene function and a property of a particular site (e.g., the prevalence of certain amino acids in buried residues in protein cores [Čerňanskỳ et al. 2014] or exposed residues such as loops [Minuchehr and Goliaei 2005]). The idea behind this research is that an expert can see which state among the inferred ML states could lead to functional differences in inferred ancestral sequences. This hybrid approach would allow us to consider the variability of ancestral inference rather than a single MPA with a degree of error or bias.

This report presents a computational approach inferring the ASR of squamates' NRL TF in selected vertebrates. The experimental work will be performed by Prof. B. W. Chang's molecular biology research team based at the University of Toronto. The laboratory work aims to use the computational results as plausible hypotheses and, in turn, allows us to validate this computational approach. We acquired and analyzed a dataset of NRL TFs sequences from 81 vertebrate species involved in developing rod photoreceptors (see Tab. 5.1 in the Appendix for the exact sequence list). Using the phylogenetic state-of-the-art method, ARPIP, we have inferred the ancestral sequence of selected vertebrates with their probability profiles [Jowkar et al. 2023].

5.2 Material and method

5.2.1 Algorithmic data acquisition pipeline

The ASR data analysis pipeline is depicted in Figure 5.1:

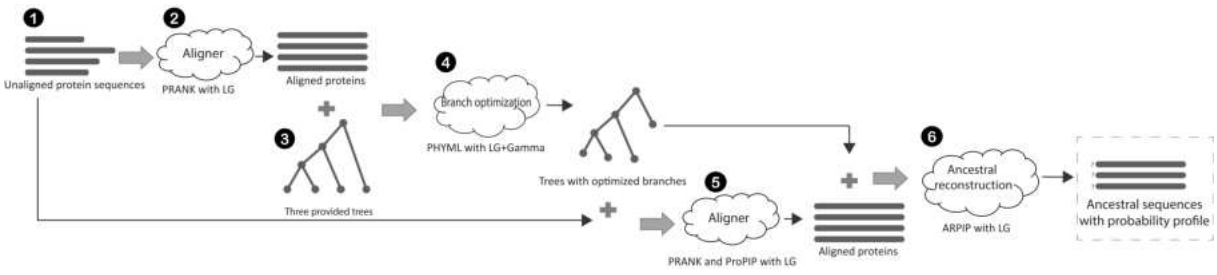


Figure 5.1: Algorithmic data acquisition pipeline (For details on each step, see section 5.2).

1. Data: a set of unaligned homologous mRNA sequences from 81 vertebrate species. These sequences represent the NRL TF involved in developing rod photoreceptors in vertebrates. The extracted RNA sequences range from a minimum length of 696 BP to a maximum

length of 865 BP, while after translation, the corresponding protein sequences range from 229 AA to 269 AA.

2. We have inferred the initial MSA using the PRANK aligner [Löytynoja 2014] with the LG substitution model [Le and Gascuel 2008] and an active insertion option (denoted as PRANK$_{+F}$). PRANK is a progressive multiple sequence aligner iteratively working on a guide tree from tips to the root, each time aligning two sister nodes while distinguishing insertions from deletions. Notice that PRANK estimates the guide tree for alignment from the raw unaligned sequences.

3. Our collaborators provided three tree topologies with priority, namely: A1-molecular squamate tree (geckos basal), in which turtles are sister to birds and crocodilians; A2-morphological squamate tree (iguanids basal), in which turtles are sister to birds and crocodilians [Gauthier et al. 2012]; B-molecular squamate tree (Geckos basal) with morphological evidence for lepidosaurs sister to turtles. Notice that the first morphological tree was constructed by taking Turtles as an outgroup, while the second used archosaurs (see Fig. 5.2 for choosing squamate ancestors in different tree topologies).

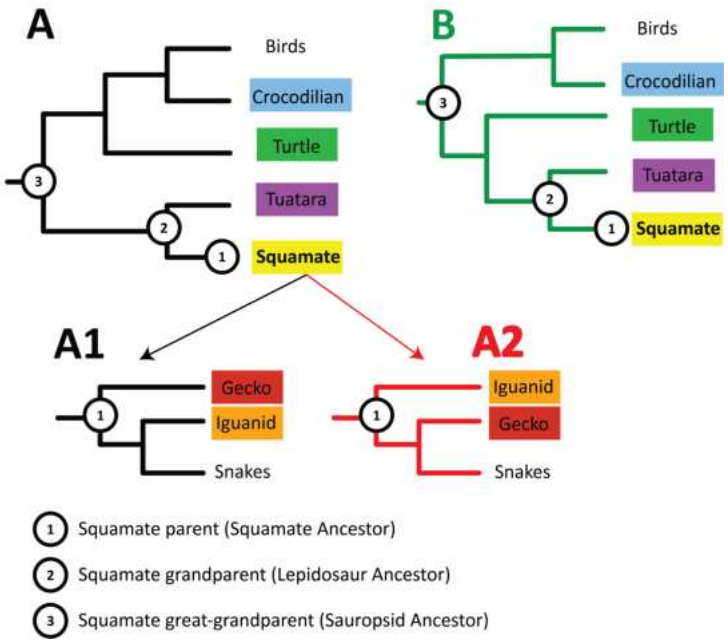


Figure 5.2: Reference trees for ASR of squamates highlighting the nodes of interest. In this project, we are interested in the immediate parent, grandparent, and great-grandparent of squamates. (Modified from the figure provided by E. Dong)

4. As an additional step, we used ML parameter optimization as implemented in PhyML [Guindon et al. 2010] with the LG+Gamma option to optimize the branch lengths of all three individual trees based on the corresponding MSAs. Notice that the tree topologies

were fixed to the input trees. However, the trees were rerooted on the branch father to amphibians (*xenopus laevis* to *microcaecilia unicolor*) so that the node order would be consistent with the considered tree topologies for interpretation (Figs. 5.8, 5.9, and 5.10 in the appendix are optimized trees under the ML and the sets of interested nodes are marked).

5. As a result of the previous step, we have three sets of MSA/tree pairs (MSA from PRANK and tree with the branch length optimized by PhyML). These tree sets, along with unaligned sequences, were provided to PRANK and ProPIP [Maiolo et al. 2021] with the LG substitution model for the final alignment inference. ProPIP also considers indels with an explicit biological model called PIP. Notice that after the translation of mRNA to AAs, the data contained three ambiguous characters in sequences from "*Phelsuma madagascariensis grandis*" and "*Rhinocheilus lecontei tessellatus*". Notice that ProPIP does not handle ambiguous characters. We removed them from unaligned sequences before ProPIP alignment, while ambiguous characters remained intact for PRANK alignment.

6. Alignment inference is also associated with much uncertainty, especially for divergent sequences, as in this case. Therefore, we performed ASR using alternative alignments, as inferred by the tree-aware alignment methods implemented in ProPIP and PRANK on three sets of trees (Figs. 5.8, 5.9 and 5.10 in Supplementary material section 5.6). As a final step, ASR using ARPIP provided us with the inferred ancestral sequences and probability profiles of states, including indel reconstruction. We used an extended LG substitution/deletion model. For the LG substitution model, we have used the option to calculate the equilibrium frequencies from data for each MSA file. All the PIP parameters were estimated using the Brent optimization method implemented in ARPIP. As ARPIP currently does not account for ambiguous characters, the three mentioned ambiguous characters in the PRANK MSA file were replaced by the most frequent characters in their respective MSA column with a majority voting mechanism prior to ASR. Notice that later, we need to determine which inferred ASR fits the data the best.

5.2.2 Preliminaries: Joint ASR under PIP

The ASR task is performed, given phylogeny tree τ and MSA matrix $\mathscr{M}$, while the former represents the evolutionary relationship and the latter the sequence homology. Let $\tau = (\mathscr{E}; \mathscr{A})$ represent the rooted binary tree where $\mathscr{E}$ is the set of extant nodes, and $\mathscr{A}$ is the set of ancestral nodes including the root Ω. Matrix $\mathscr{M}$ consists of character states $i \in \Sigma$, where alphabet Σ can be the set of nucleotides or AAs.

ARPIP computes the joint ML ASR by considering the ancestral state of all the ancestral nodes

$\mathscr{A}$. This contrasts with the marginal reconstruction, where only a single ancestral node is of interest. Note that joint and marginal reconstructions do not necessarily yield the same results. Here, the set of all ancestral nodes is reconstructed while for the sake of interpretation, the set $\mathscr{A}$ is limited to nodes $\mathscr{S} \in \mathscr{A}$ denoted by numbers $\mathscr{S} = \{1,2,3\}$ on the tree topologies in Figure 5.2.

ARPIP computes the joint ASR in linear time with respect to the sequence length, assuming site independence. ARPIP consists of two main algorithms: Most likely indel point (MLIP) extraction and joint ASR. The first algorithm prunes the tree by inferring the insertion and deletion points on the tree topology for each site. The second algorithm, joint ASR, reconstructs ancestral characters on the pruned subtrees τ_{MLIP}. A brief description of the joint ASR algorithm implemented by ARPIP is presented below. More details can be found in the ARPIP manuscript [Jowkar et al. 2023].

ARPIP infers the MPA per site using the dynamic programming technique. As a first step of the joint ASR algorithm, ARPIP traverses τ_{MLIP} from $\mathscr{E}$ towards the root Ω for each node n; while traversing, it computes the conditional probabilities $L_n(i)$ and their corresponding character states $C_n(i)$ for each AA i. Usually, this conditional probability is referred to as "likelihood values" $L_n(i)$ in the literature. Notably, the likelihood $L_n(i)$ is the probability of the best reconstruction of the subtree rooted at node n, conditioned on the father of node n being assigned character state i. Therefore, character state $C_n(i)$ is the character state assigned to node n in the optimal condition reconstruction. For a better understanding, refer to the example shown in Figure 5.3.

In the next step, ARPIP traverses the tree τ top-down, picking the locally optimal solution to the reconstruction sub-problems, which provides us with the global best reconstruction at the root. It is important to note that the locally best reconstructions at internal node n are necessarily consistent with the global best reconstruction by ARPIP. For example, consider node $n = 2$ in Figure 5.3.a. In such a case, P has the maximum local likelihood value, assuming the father node $n = 1$ also has the same value. While in reality, we do not know the reconstruction of the rest of the tree. If we assume the father node $n = 3$ is assigned character state A, then the best reconstruction of node 2 is A (see Fig. 5.3.b). In this case $L_3(A) = 2.92 \times 10^{-10}$ and $C_3(A) = A$. In a similar approach, $L_2(A) = 4.24 \times 10^{-7}$ and $C_2(A) = A$. After calculating likelihood and its corresponding character state similar to the abovementioned example, the most likely reconstruction A is assigned to the root as it gives us the global best reconstruction. Then, the most likely reconstruction of the whole tree is assigned as presented in Figure 5.3.b. It is important to mention this global best reconstruction (i.e., A at the root, then the rest of the reconstruction as shown in Fig. 5.3.b) is not necessarily the global optimum reconstruction.

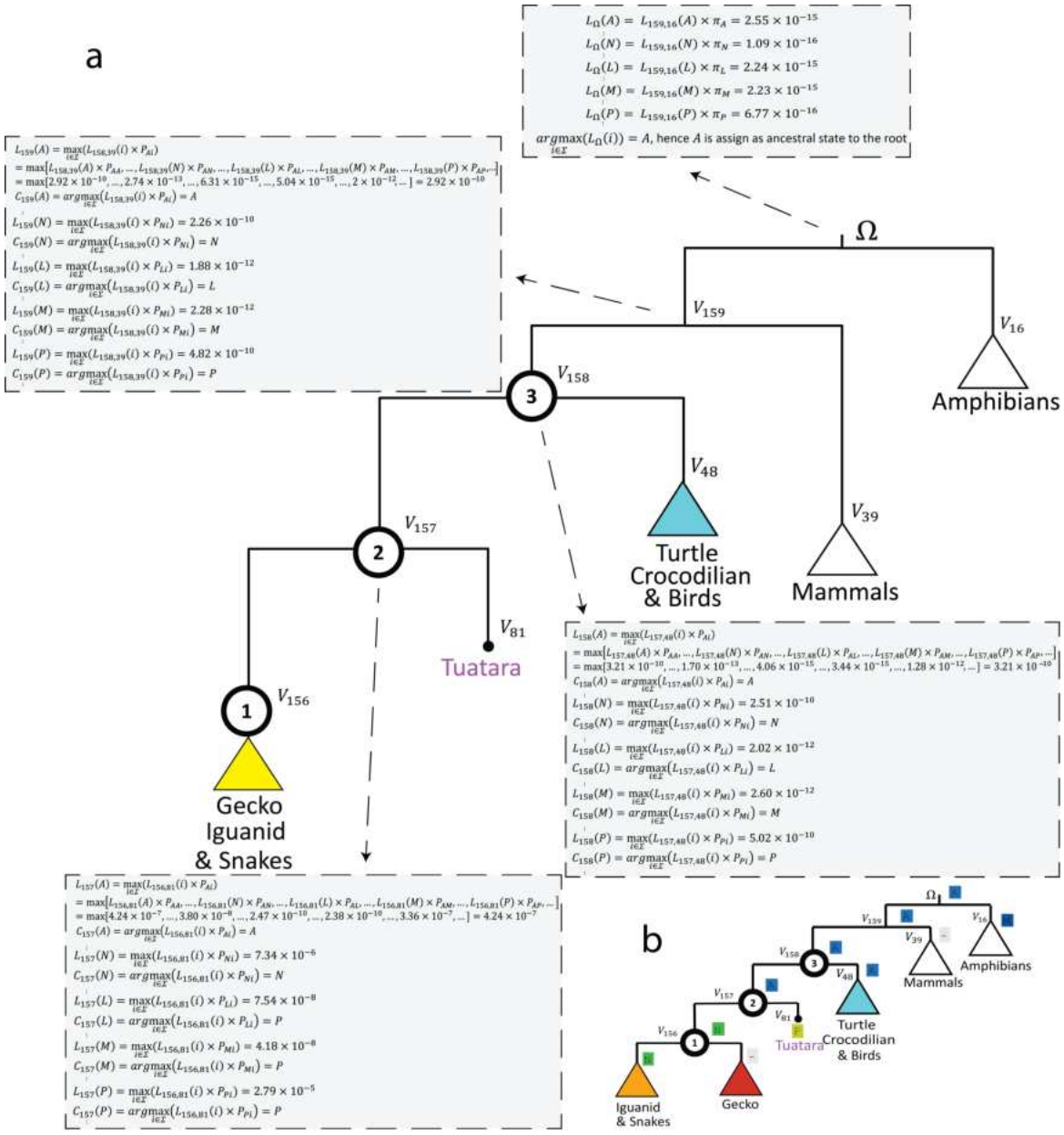


Figure 5.3: A detailed computation of ARPIP's joint ASR on nodes of interest for site 0 of PRANK alignment and tree A1. **a)** A bottom-up traversal of the tree by the ARPIP algorithm computing the conditional probabilities is presented in dashed-shaded boxes. We start by assigning A to the root Ω. Since A is assigned to root and $C_{V_{159}}(A) = A$, we assign A to node V_{159}. Since A is assigned to node V_{159} and $C_{V_{158}}(A) = A$, we assign A to node V_{158} and so forth. **b)** The reconstructed MPA character states given the provided computations.

5.2.3 A solution to ambiguity of joint ASR under PIP

The ASR tools that infer the MPA do not account for alignment and tree inference biases, additionally contributing to uncertainty in ancestral sequence inference. ARPIP, like other ML methods for ASR, provides us with the MPA character states by considering indel events. The latest version of ARPIP informs the user about the ambiguity introduced by the uncertainties in input and the ASR inference. This is done by providing the MPA and probabilities profile describing the possible character states and indel events at every sequence position.

In a joint ASR, the probability profiles reflect the conditional probability of each ancestral state conditioned on their immediate parent having an assigned specific character state [Pupko et al. 2000]. Here, we present two methods for computing the probability profile for each reconstructed MPA, namely *normalized* (Norm) and *naive posterior* (NP) probability profiles. In the first approach, per each node n, we compute the normalized conditional probability $Pr_n^{Norm}(i)$ as follows:

$$Pr_n^{Norm}(i) = \frac{L_n(i)}{\sum_{\sigma \in \Sigma} L_n(\sigma)} \tag{5.1}$$

Where $Pr_n^{Nrom}(i)$ is the Norm probability for state i at node n and $L_n(i)$ represents the likelihood value of character state i for node n.

The second approach uses background frequencies π_i computed directly from the data to calculate the probability profile of each character state i per site. We consider the equilibrium frequency π as a prior knowledge of AA frequencies. This NP probability profile reflects the confidence degree of each ancestral state for each character i and node n per site. The computation of $Pr_n^{NP}(i)$ comes as follows:

$$Pr_n^{NP}(i) = \frac{L_n(i) \times \pi_i}{\sum_{\sigma \in \Sigma} L_n(\sigma) \times \pi_\sigma} \tag{5.2}$$

Where π_i is the equilibrium frequency of character state i computed directly from the data under the specified substitution model (LG in this case).

5.3 Results

5.3.1 Squamates reconstructed ancestral sequences

The dataset used in this report consists of 81 vertebrates NRL TF sequences aiming to infer the ancestor of squamates. We have performed ASR on two sets of alignments (inferred by tree-aware ProPIP and PRANK) based on three sets of trees (A1, A2, and B). This gave us six sets of reconstructed ancestral sequences. We used the two mentioned aligners to obtain ancestral sequences, as they treat indels differently. PRANK accounts for long indels, while ProPIP models indels explicitly but assumes only one-residue indel. We used three tree topologies to see the variability of inference with respect to differences in tree out-grouping mechanisms provided to the estimation. We focused on three nodes of interest in each set: squamate parent (node 1), squamate grandparent (node 2), and squamate great-grandparent (node 3).

To account for the inference variability of each squamate ancestor, different sets were used to infer their ancestral sequences. The reconstructed ancestral states for nodes of interest across different tree topologies are shown in Figure 5.4. The highlighted regions illustrate the variability of reconstruction due to the tree parameters, which allows the expert to find the set that best matches their criteria. For the remainder of this section, we would only consider the ASR results based on the A1 tree (see Fig. 5.8 in the supplementary materials) and the PRANK alignment to illustrate the results.

5.3.2 Interpreting uncertainties in ancestral state inferences with probability profiles

In this section, we explain and illustrate the information represented by the MPA and the probability profile using the A1 tree and PRANK alignment. Figure 5.5 shows logo plots, reflecting the probability profile of specified ancestral nodes and the final MPA inference for the joint ASR. In some positions (Fig. 5.5.a), AA selected as the MPA has a high probability reflecting the same value, while in some cases, the MPA and the probability profile hold different values (Fig. 5.5.b). This behavior can be further explained by joint conditional probability computation. When determining the MPA in the joint reconstruction of ancestral nodes, we consider all the information from both descendants and ancestors to make the decision. In comparison, the probability profiles reflect the local decisions only from descendants in an early step of the computation.

In ARPIP's joint likelihood computation, we assume that character state i is assigned to the father of node n. This means that the likelihood value is given based on the value assigned to

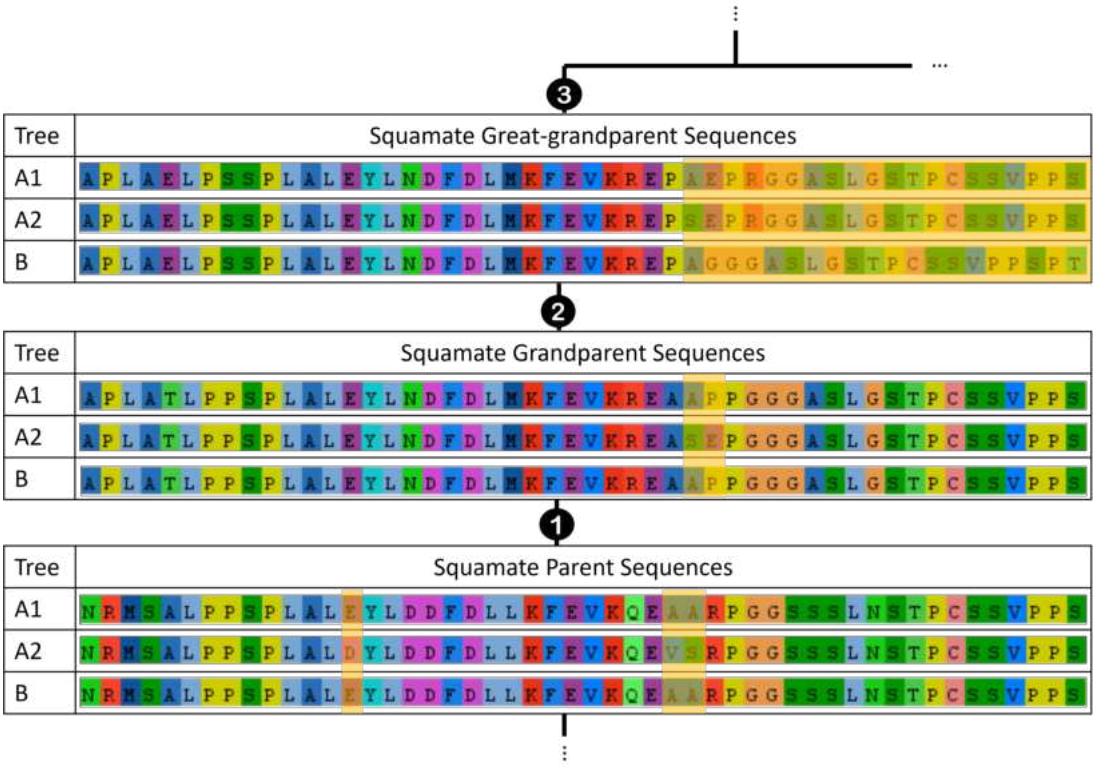


Figure 5.4: Sites 0-50 of ARPIP's ancestral sequences across three mentioned tree topologies from PRANK alignment. The highlighted sections are diverse due to the differences in tree topology structure and branch lengths.

the father. For example, in the case of site 0, if the father is assigned character state A, then we pick $C_\Omega(A)$, which is A. Further, when computing the likelihood $L_n(.)$, only the immediate father of a node n is considered. The best reconstruction for a subtree supported by that node is independent of the reconstruction of the rest of the tree. This means that in the computation of $L_n(.)$ and $C_n(.)$, no information from the ancestral node other than the immediate father is considered. Since $L_n(i)$ is the likelihood value of the best reconstruction of this subtree conditioned on that the father of node n is assigned character state i and $C_n(i)$ is its corresponding character state. Instead, we consider all the inferred information on the tree to decide on the joint reconstruction.

The numerical example in Figure 5.3 demonstrates the ASR computation. Here, at site 0 for node 2, we observe an uncertainty in the MPA as the amino acid P has the maximum conditional probability $L_2(P)$ while A is assigned as MPA. As explained earlier, this is rooted in the conditional probability computation as it only reflects information of subtree rooted at node 2. It assumes that node 3 (the father of node 2) has character state A as MPA. This implies that if node 3 had another character state assigned as the ancestral state (L, for example), the inferred MPA yields a different value ($C_2(L) = P$). This would result in having P as the MPA of node 2.

Notice that having A as the MPA at node 2 provides us the ML value (global optimum), while picking another value is not the optimal solution. Biological experts can use additional prior information, such as functional properties of specific proteins and positions, since such biological knowledge can help to "overwrite" the uncertainty inference due to various biases introduced via the input MSAs, tree, and the model itself. In reality, it is different as A has been assigned to node 3. Therefore, $L_2(A)$ and $C_2(A)$ (which in this case is A) are assigned as an ancestral state of node 2.

Two probabilistic methods were introduced earlier to handle uncertainty in the inference. Figure 5.6 provides a comparison of these alternative solutions. We believe that the normalized conditional probability (Norm) is the most coherent one. Let us consider the case where both profiles do not match the MPA. Assuming that the choice of substitution/indel model (in this case, extended LG model) is correct. We know the substitution model may introduce a bias toward the more frequent AA state (in this case A and L) to our inference [Pollock and Chang 2007]. This bias is magnified by multiplying the background frequency by raw conditional probabilities to compute the NP conditional probabilities. For example, Figure 5.6 highlighted regions where the MPA and profile are inconsistent; the NP exaggerates the signal toward the more frequent AA. In Figure 5.6.a using NP, the ASR over magnifies the probability profile as we multiply the probability with background frequency π. Multiplying $L_n(.)$ with π at the root appears incongruous as we already have such a factor in the conditional probability. However, this is sensible for other internal nodes. It is important to mention that assuming a prior bias toward more frequent AA, the NP could sometimes be misleading as we add another bias by magnifying the background frequency. In particular, it is crucial when the NP signal disagrees with MPA; it could amplify (Fig. 5.6.b) or weaken (Fig. 5.6.c) the signal toward MPA.

5.4 Discussion

Ignoring uncertainty in the ASR is a serious limitation of current ASR implementations, resulting in overconfidence of ASR point estimates [Oliva et al. 2019]. MPAs are known to be biased toward the model parameter from substitution model, tree, and MSA. It results in cumulative biases through the inference procedure. For instance, assuming the correct substitution model, we are prone to bias toward more frequent AA during the ASR computation. Different techniques have been proposed in the literature to reduce these uncertainties in each inference step before ASR [Gaucher et al. 2008, Oliva et al. 2019, Pollock and Chang 2007], which are beyond the scope of this research.

Observing other character states in internal nodes is possible during the conditional probability computation but with a lower chance [Pupko et al. 2000]. In our experiment, we observed this

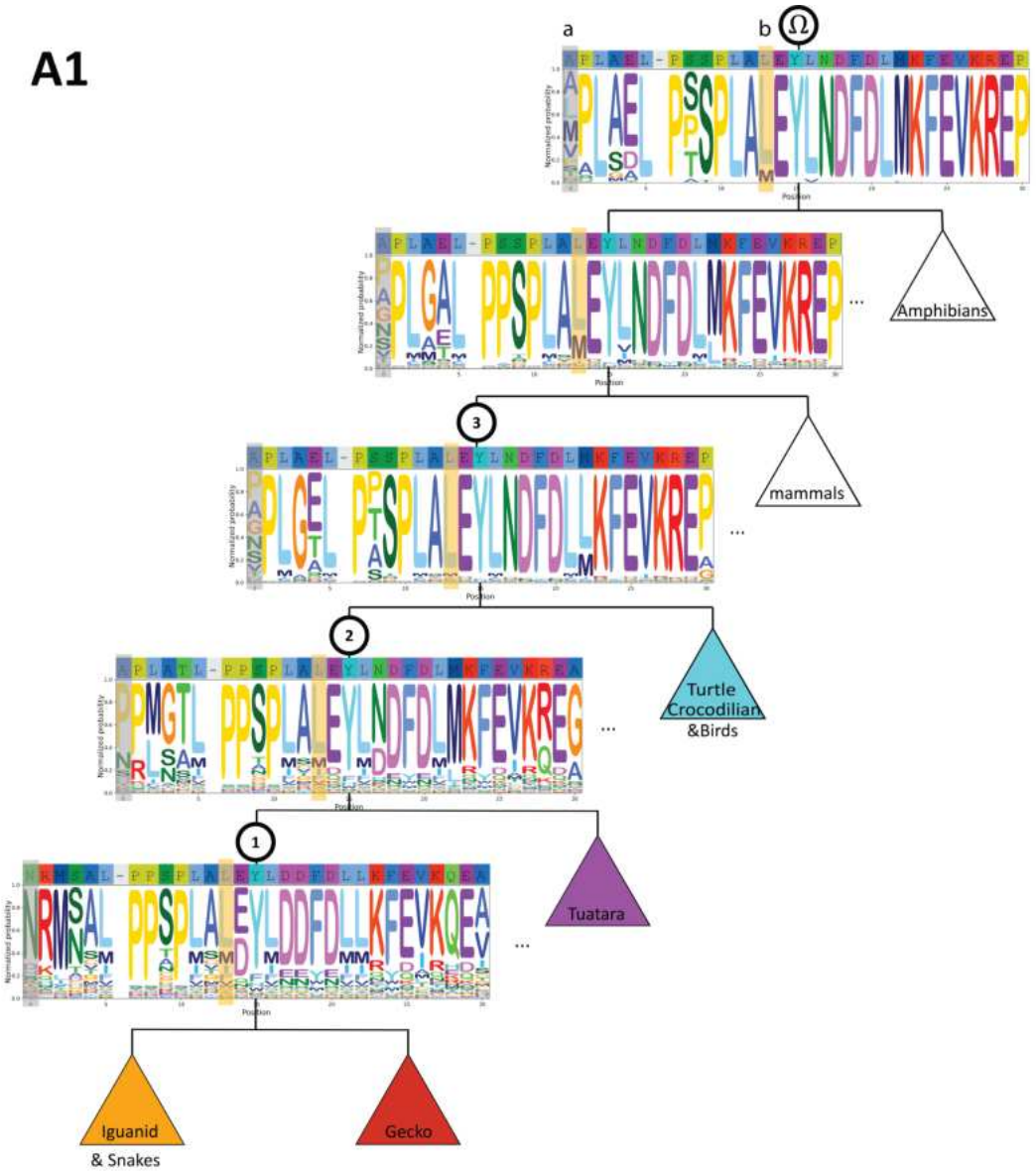


Figure 5.5: An example of the ARPIP probability profile. The top row of each node shows the inferred ancestral sequence, while the second row is the conditional probability profile of the corresponding site. The three selected nodes of interest augmented with the root reflect the joint ASR. Highlight regions differ in reconstruction sites 0 and 13 in the probability profile and assigned ancestor of the squamate's ancestors. **a)** Grey region, in which the MPA and the probability profile are inconsistent, and **b)** golden region, in which the MPA and the probability profile match.

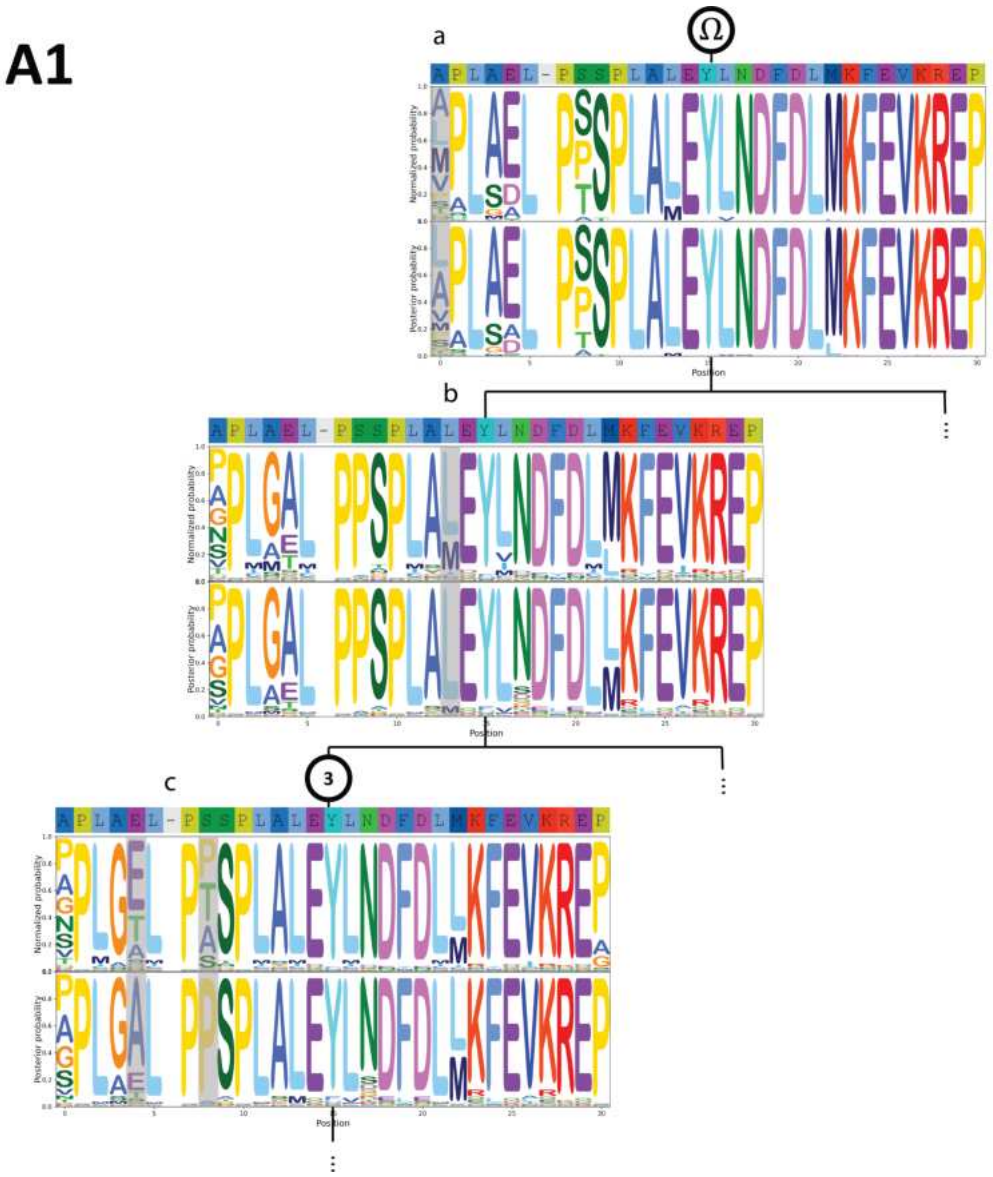


Figure 5.6: An example comparing Norm and NP probability profiles. The first row shows the MPA, while the second and third rows reflect Norm and NP probabilities. For site 0: **a)**, the NP probability magnifies the background frequency signal where the profile disagrees with the MPA. When the probability profile and the MPA disagree **b)** NP amplifies **c)** NP weakens the signals towards the MPA.

artifact as presented in Figure 5.5.a. In the case of Figure 5.5.a, the bias was inherited from the root node Ω due to the higher background frequency of AA A over M in our dataset. As a result, AAs with less frequency in the data have a lower chance of substitution. This is one of the thousands of cases in which the frequency of AA states changed the inference, assuming the choice substitution model and other parameters are correct. Here, we only investigate some theoretical aspects of the problem and present a computational approach when the goal is to reconstruct ancestral sequences in the laboratory.

Generally, it is established that reconstructions are biased toward model parameters, including the most frequent AAs [Pollock and Chang 2007]. The background (equilibrium) AA frequency (π) is usually computed from target sequences or large databases. To reduce the bias to the most frequent AA, it is suggested to compute it from a large database. In this study, we have computed this parameter from the relatively small target data. We believe this inference has some bias as the target study is relatively small while the sequences are separated by large evolutionary distances. One possible solution is to do the site-specific reconstruction while computing the background frequency of a reconstruction site, i.e., using site AA frequency profiles such as in the CAT model by [Lartillot and Philippe 2004]. In other words, we infer the ancestral states with site-specific background frequencies instead of the global ones that are computed over the whole MSA length.

5.4.1 Ambiguity of ASR inference

Due to the statistical uncertainty, we have proposed accounting for this ambiguity when using the MPA, by considering the presence of multiple possibilities in reconstructing those sequences and additional biological knowledge and interpretations. Some uncertainties can impact the reconstruction through affecting the ancestral protein's functional features. The probability profile solution for ASR helps to unveil relevant information about the inference uncertainty to biology experts. Most importantly, this approach has no computational overload, as all the computation is done in the early stages. We thus recommend that these probability profiles be considered alongside the MPA.

Traditional approaches to handling ambiguity of inference result in several potential sources of error. In cases where the reconstruction does not provide a clear-cut confident solution, inferring the biochemical and structural properties of ancient genetic sequences could be helpful. The issue of uncertainty in the experimental recreation of ancestral sequences arises, which we tried to address by introducing a probability profile. The question still remains of whether the functional features of the reconstructed protein are a reliable approximation of the functional features of the ancestral proteins [Pollock and Chang 2007]. The practical niche to address the ambiguity of reconstruction is to consider alternative reconstruction and consider the variability

of reconstruction.

With the help of probability profiles, one can obtain the result in the format that "at node n, the character state i has the probability of p_i while the character state j has the probability p_j". During the joint reconstruction, the ASR algorithm picks the character with the best global reconstruction and assigns j instead of i although $p_i > p_j$ due to the high global likelihood value. If $p_i > p_j$, then one has this option to pick character i over j concerning the functional consequence. Figure 5.7 represents such an alternative reconstruction. Clearly, this alternative reconstruction is not the global optimum solution to the reconstruction; however, experts could decide on the necessity of such an alternative reconstruction based on the functional consequences of the ancestor. We have included a Python script to supplement this alternative reconstruction for a specific subtree.

Along with the ambiguity of reconstruction, we have tried to consider the variation of the reconstructed solution. Indeed, we proposed selecting one of the multiple best reconstructions with the help of alternative MSAs and trees. This allows to consider the sensitivity to assumed alignment and phylogeny as reflected by the extent of variability of the reconstruction. The input MAS and tree may be inferred with a set of specific assumptions, which should be coherent with the model used for the ASR, particularly with respect to modeling indel events. For this reason here we used aligners PRANK and ProPIP that account for indel evolution rather than solely relying on gap penalties. The choice of aligner could cause different ancestral reconstructions due to their handling of insertion and deletion events. Further, the aligner typically contributes to the tree inference, potentially introducing additional bias to the ASR inference. Indeed, MSAs are later used to optimize trees and compute substitution model parameters. Ideally, joint inference of MSA and trees should be considered where possible. Currently, only a Bayesian approach can be used for this task [Redelings 2021], but it is feasible only for relatively small datasets.

5.5 Concluding remarks

The main motivation behind this research is to improve how the statistical uncertainty in ASR inference is dealt with. The issue of uncertainty in the experimental recreation of ancestral protein arises, which we tried to address by introducing a probability profile. In the case that multiple AAs have slightly different probability profiles, selecting one of them might be misleading as the values might be different due to biases introduced by the substitution model, tree, or MSA. These values affect the conditional probabilities computed by the ASR tool and the joint reconstruction.

Photoreceptors have been the subject of many studies, but we are still far from understanding

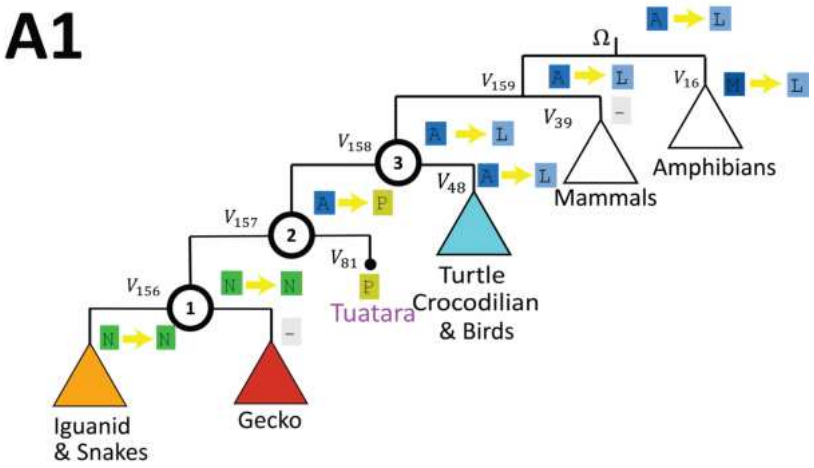


Figure 5.7: The choice of the ancestral state for the root will play a major role in the joint reconstruction of other nodes. In this example, we illustrate the effect of such a choice. If a biologist decides that the choice of AA *A* at the root for site 1 of MASA should be *L*, then the whole reconstruction would be affected. We assign *L* to the node Ω.

the molecular evolutionary trajectory of photoreceptor cells and relevant proteins. The NRL protein was chosen for ASR partly because of its key role in photoreceptor development in vertebrates. NRL of the photoreceptor is highly important as the choice of a cell to be rod or not is largely determined by this TF [Swaroop et al. 2010]. The transmutation theory explains how vertebrate photoreceptors transform from one kind to another [Schott et al. 2016, Walls 1934] while there is some evidence that rods are most likely evolved from cons [Simoes et al. 2016]. With respect to the evolutionary transmutation of photoreceptors, some studies were performed only on a small group of vertebrates (snakes and lizards)[Simoes et al. 2016, Schott et al. 2016, Walls 1934].

It is believed that squamate retinas are more evolutionarily dynamic than those of fishes, birds, or mammals [Simoes et al. 2016]. Therefore, here we have investigated the substitution and indel mutations in NRL responsible for the transmutation of photoreceptors in squamate ancestors. We proposed a hybrid solution to handle this uncertainty by simultaneously considering MPAs and probability profiles.

On the other hand, recreating the ancestral proteins in the laboratory can help us to discover the functional differences between the ancestral proteins. Based on squamate ancestral inferences, the most likely protein sequences can be synthesized, expressed, and assayed for functional analysis in the laboratory.

The idea behind this study is to see which molecular changes in the reconstructed evolutionary trajectory could have led to functional differences between the ancestral and present-day

proteins. Here, we consider the variability of ancestral inference rather than point estimates of the MPA [Pollock and Chang 2007]. Considering alternative reconstructions helps to validate different hypotheses about the evolution of cone retina cells in squamates, including the transmutation. Laboratory experiments can illustrate how ancestral dynamics can be compensated by the evolutionary modification of the existing cellular structures [Schott et al. 2016].

5.6 Contributors

The conceived and designed experiments, including the data and the case study scenario questions, were provided by Prof. B.W. Chang and E. Dong, while Gh. Jowkar performed the computational experiments. All the authors contributed to the data analysis and interpreting the results. Gh. Jowkar wrote this technical report.

Appendix

Supplemental materials

Selected sequences

Name	Organism	Group	Accession Number
M_unicolor_NRL	Microcaecilia unicolor	Amphibian	XM_030187778.1
R_bivittatum_NRL	Rhinatrema bivittatum	Amphibian	XM_029580581.1
X_laevis_NRL	Xenopus laevis	Amphibian	NM_001085737.1
X_tropicalis_NRL	Xenopus tropicalis	Amphibian	XM_012953266.2
Geotrypetes seraphini	Geotrypetes seraphini	Amphibian	XM_033923446.1
Bufo bufo NRL	Bufo bufo	Amphibian	XM_040416857.1
Leptobrachium leishanese NRL	Leptobrachium leishanese	Amphibian	ENSLLET00000013251.1
Nanorana parkeri NRL	Nanorana parkeri	Amphibian	XM_018556114.1
Rana temporaria NRL	Rana temporaria	Amphibian	XM_040354086.1
A_mississippiensis_NRL	Alligator mississippiensis	Crocodilian	XM_014526180.2
C_porcellus_NRL	Cavia porcellus	Mammal	NM_001173050.1 2
H_sapiens_NRL	Homo sapiens	Mammal	NM_006177.5
M_musculus_NRL	Mus musculus	Mammal	NM_008736.3
O_anatinus_NRL	Ornithorhynchus anatinus	Mammal	XM_029078437.1
O_cuniculus_NRL	Oryctolagus cuniculus	Mammal	XM_002718089.3
O_garnettii_NRL	Otolemur garnettii	Mammal	XM_012812806.2
P_troglodytes_NRL	Pan troglodytes	Mammal	XM_001166004.5
Phascolarctos cinereus NRL	Phascolarctos cinereus	Mammal	XM_020984450.1
R_norvegicus_NRL	Rattus norvegicus	Mammal	NM_001106036.2 2
T_manatus_NRL	Trichechus manatus	Mammal	XM_004390499.1
Vombatus ursinus	Vombatus ursinus	Mammal	XM_027842962.1
Ursus arctos horribilis	Ursus arctos horribilis	Mammal	XM_026486438.2
Ophisaurus gracilis NRL	Ophisaurus gracilis	Squamate	Ogr03986
Shinisaurus crocodilurus NRL	Shinisaurus crocodilurus	Squamate	ENSACAP00000015884-D1
Varanus komodoensis NRL	Varanus komodoensis	Squamate	Komodolmaker-scaffold112-snap-gene-7.10-mRNA-1 - ORF 1 (frame 1)
Arizona elegans NRL	Arizona elegans	Squamate	c19651_g1_i1 - ORF 17 (frame 1)
Cemophora coccinea NRL	Cemophora coccinea	Squamate	
Hypsiglena torquata NRL	Hypsiglena torquata	Squamate	TR39015|c0_g1_i1 - ORF 1 (frame 1)
Lampropeltis getula floridana NRL	Lampropeltis getula floridana	Squamate	
Masticophis flagellum NRL	Masticophis flagellum	Squamate	
Pantherophis guttatus NRL	Pantherophis guttatus	Squamate	
Pantherophis obsoletus NRL	Pantherophis obsoletus	Squamate	WJSR01000377.1
Phyllorhynchus decurtatus NRL	Phyllorhynchus decurtatus	Squamate	

Rhinocheilus lecontei tessellatus NRL	Rhinocheilus lecontei tessellatus	Squamate	
Thamnophis sirtalis NRL	Thamnophis sirtalis	Squamate	XM_014070535.1
Thermophis baileyi NRL	Thermophis baileyi	Squamate	QLTV01001795.1
Hydrdophis hardwickii NRL	Hydrdophis hardwickii	Squamate	RSAD01054254.1
Hydrophis cyanocinctus NRL	Hydrophis cyanocinctus	Squamate	RSAE01495440.1
Hydrophis melanocephalus NRL	Hydrophis melanocephalus	Squamate	BHFS01108563.1
Laticauda colubrina NRL	Laticauda colubrina	Squamate	BHFR01000527.1
Laticauda laticaudata NRL	Laticauda laticaudata	Squamate	BHFT01016592.1
Naja naja NRL	Naja naja	Squamate	CM019155.1
Notechis scutatus NRL	Notechis scutatus	Squamate	XM_026683089.1
Ophiophagus hannah NRL	Ophiophagus hannah	Squamate	AZIM01001134.1
Pseudonaja textilis NRL	Pseudonaja textilis	Squamate	XM_026719572.1
Ptyas mucosa NRL	Ptyas mucosa	Squamate	WNWU01000310.1
Correlophus ciliatus NRL	Correlophus ciliatus	Squamate	Correlophus.....
Eublepharis macularius NRL	Eublepharis macularius	Squamate	CCG023294.1
Gekko gecko NRL	Gekko gecko	Squamate	
Paroedura picta NRL	Paroedura picta	Squamate	Parpi0017749.t1
Gehyra mutilata NRL	Gehyra mutilata	Squamate	Gehyra_CL1Contig954_...
Gekko japonicus NRL	Gekko japonicus	Squamate	XM_015413654.1
Hemidactylus turcius NRL	Hemidactylus turcius	Squamate	Hemidactylus
Lialis burtonis NRL	Lialis burtonis	Squamate	Lialis_TRINITY
Phelsuma laticauda NRL	Phelsuma laticauda	Squamate	Phelsuma_CL4539Contig1_1
Lepidodactylus lugubris NRL	Lepidodactylus lugubris	Squamate	Lepidodactylus_CL5798Contig1_2
Phelsuma madagascariensis grandis NRL	Phelsuma madagascariensis grandis	Squamate	
Anolis carolinensis NRL	Anolis carolinensis	Squamate	XM_008124710.2
Chamaeleo calyptratus NRL	Chamaeleo calyptratus	Squamate	TRINITY_DN41998_c4_g1_i3_1 - ORF 8 (frame 2)
Pogona vitticeps NRL	Pogona vitticeps	Squamate	XM_020809326.1
Sceloporus undulatus NRL	Sceloporus undulatus	Squamate	XM_042476348.1
Lacerta agilis NRL	Lacerta agilis	Squamate	XM_033170397.1:52-966
Lacerta bilineata NRL	Lacerta bilineata	Squamate	Lbil_239
Lacerta viridis NRL	Lacerta viridis	Squamate	Lvir_2172
Podarcis muralis NRL	Podarcis muralis	Squamate	XM_028703022.1
Salvator merianae NRL	Salvator merianae	Squamate	ENSSMRT00000005745.1
Tupinambis teguixin NRL	Tupinambis teguixin	Squamate	RS eye transcriptome
Zootoca vivipara NRL	Zootoca vivipara	Squamate	XM_035137209.1
Boa constrictor NRL	Boa constrictor	Squamate	BoaCon10541-RA
Python bivittatus NRL	Python bivittatus	Squamate	XM_007437307.3
Crotalus horridus NRL	Crotalus horridus	Squamate	LVCR01046683.1
Crotalus tigris NRL	Crotalus tigris	Squamate	XM_039338079.1
Deinagkistrodon acutus NRL	Deinagkistrodon acutus	Squamate	Dacu_08853

Protobothrops mucrosquamatus NRL	Protobothrops mucrosquamatus	Squamate	XM_029285314.1
Vipera berus berus NRL	Vipera berus berus	Squamate	KN632707.1
Sphenodon punctatus NRL	Sphenodon punctatus	Tuatara	ScrUdWx_1027
Emydocephalus ijimae NRL	Emydocephalus ijimae	Turtle	BHEV01135681.1
Chelonia mydas	Chelonia mydas	Turtle	XM_037877092.2
Dermochelys coriacea NRL	Dermochelys coriacea	Turtle	XM_038370154.2
Gopherus evgoodei	Gopherus evgoodei	Turtle	XM_030542752.1
Mauremys reevesii NRL X	Mauremys reevesii	Turtle	M_039498308.1:272-994

Table 5.1: The sequence information of selected 81 species.

Three tree topologies after branch length optimization

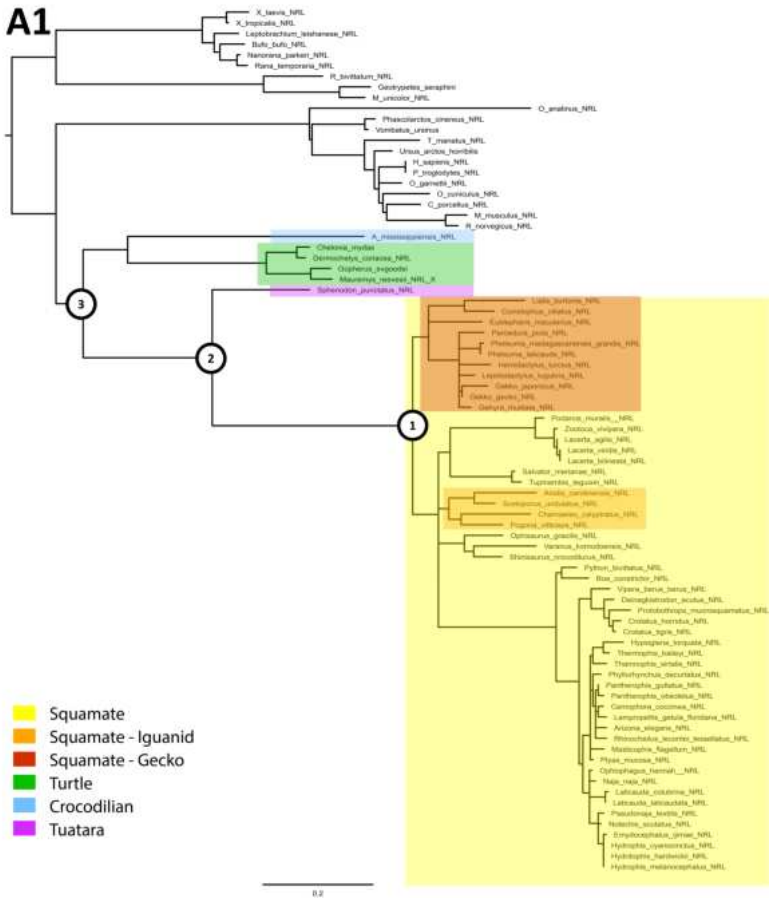


Figure 5.8: Molecular squamate tree in which turtles are sister to birds and crocodilians with highlighted nodes of interest. (Modified from the figure provided by E. Dong)

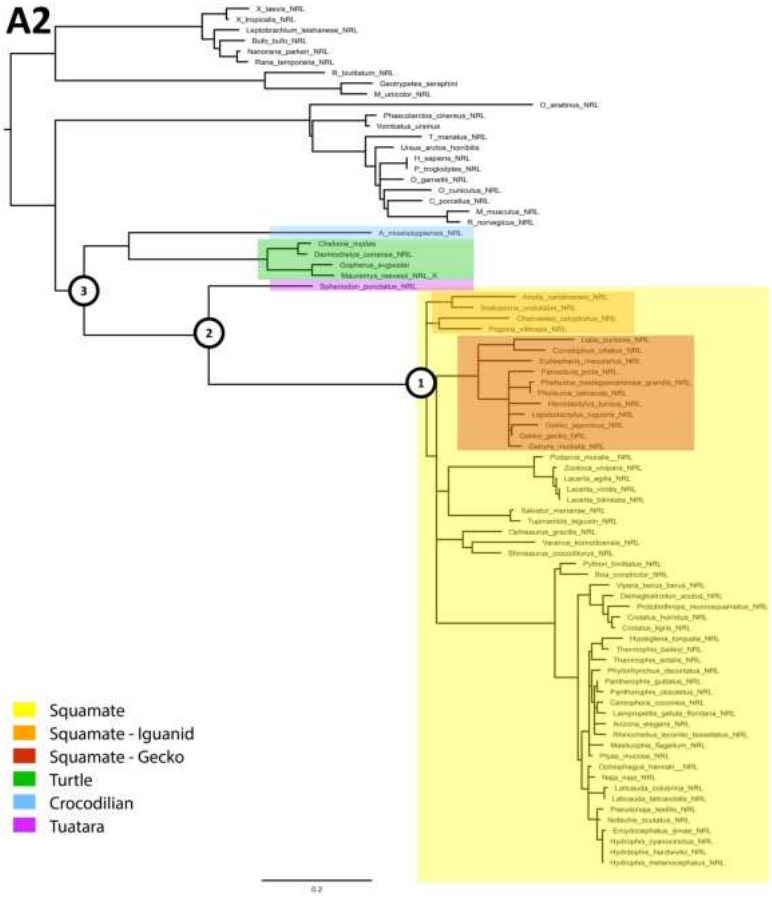


Figure 5.9: Molecular squamate tree in which turtles are sister to birds and crocodilians with highlighted nodes of interest. (Modified from the figure provided by E. Dong)

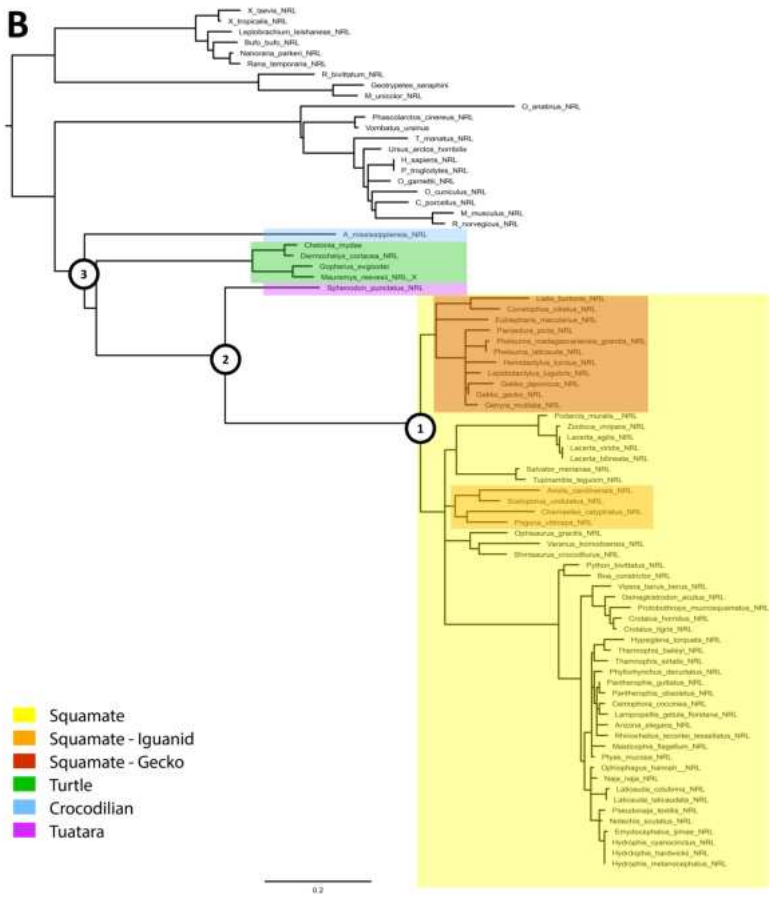


Figure 5.10: molecular squamate tree with morphological evidence for lepidosaurs sister to turtles. (Modified from the figure provided by E. Dong)

Chapter 6

Conclusions and future works

Phylogenetic studies are motivated by various scientific queries, from questions in evolutionary biology to protein engineering. Reconstruction of the evolutionary history of molecules is a great challenge for researchers, both from statistical and computational perspectives. In this thesis, the question of ASR has been studied from the theoretical and computational viewpoint. This concluding chapter will briefly summarize the contributions of this thesis while highlighting the strengths and limitations of the current research and will propose new directions for future studies.

6.1 Summary of thesis

The reconstruction of ancestral sequences from contemporary data is a powerful tool for understanding molecular evolution due to mutation. Existing ASR tools often ignore indels, while others fail to consider a sensible model of indels. Moreover, most existing ASR frameworks provide only MPA, which maximizes the probability of ancestral sequences given the present-day data and model parameters. Their strategy fails to handle the uncertainty associated with ASR inference. This research addressed the mentioned limitations of ASR by focusing on an adequate model of indel and treating/studying inference uncertainty.

This thesis introduced ARPIP by reconstructing the ancestral sequence of spike proteins of the beta family of SARS-COV-2 (chapter 3). ARPIP combines a popular ASR technique [Pupko et al. 2000] with the PIP model [Bouchard-Côté and Jordan 2013] of character substitutions and indels. This approach enables the reconstruction of ancestral sequences more accurately, including evolutionarily significant indels, and to identify the most probable indel events on a given phylogenetic tree. In the PIP model, the treatment of gaps is computationally more efficient, biologically more realistic, and phylogenetically more reasonable than existing models.

Then, the current thesis presented an investigation of indel patterns in high-confidence orthologous proteins of six mammalian and simulated data studying the PIP model (chapter 4). PIP models long indels as several independent single-site events, keeping the gap length distribution of ancestral lineage-given descendants. This thesis demonstrated that the PIP, a single-character indel model, can preserve the gap distribution in the ASR task. Moreover, this study implemented an exploratory analysis to investigate the correlation between indel and the phylogeny dynamics.

Finally, this thesis proposed two types of probability profiles to reflect the MPA's uncertainty, illustrated on 81 vertebrates' NRL TF (chapter 5). The MPA inferred by the ASR is an estimation of a true ancestral sequence. ASR involves multiple interconnected phylogeny tasks that introduce bias to the inference as each step carries uncertainty signals. ARPIP, like some other programs (e.g., [Ashkenazy et al. 2012]), provides the MPA and the probability profile reflecting the posterior probability of inferred ancestral sequence as confidence at each inferred ancestral character.

It is known that ASR methods vary due to the strategy they are implementing to address some of the above aspects of ASR with respect to the existing limitations. This research has developed, implemented, and evaluated a unique ASR method in the frequentist framework (i.e., EB approach) capable of modeling both short and long indels in linear time. This thesis has provided a more realistic model of indel mutation in ASR as the gaps are mapped on the phylogeny under

the PIP model. ARPIP is accompanied by posterior probability profiles to reflect the uncertainty of MPA sequences.

6.2 Practical limitations and improvements

From a methodological viewpoint, different studies investigate the limitations of ASR methods (e.g.,[Vialle et al. 2018, Selberg et al. 2021]). Like other phylogenetic tasks, ASR is prone to technical and theoretical limitations. This thesis addressed one of the biggest ASR limitations by adapting PIP to model substitution and indel. However, we need a realistic model of indel with rate variation as different sites evolve at different rates.

Another ongoing research subject of ASR is the potential functional bias of the likelihood estimation[Selberg et al. 2021]. To assess whether an MPA is accurate, ASR methods should design a metric to reduce a bias in the inference or develop methods to fit the real-world ancestral data. The former is done in various studies (e.g., Oliva et al. [2019]), while the latter is a big constraint of ASR tasks as no big data of ancient species is available. That makes the validation process of molecular evolution (inferences about the past based primarily on examining present-day protein information) extremely difficult [Pollock and Chang 2007]. The rest of this section will address some of the ASR methods' impediments with their possible solutions:

6.2.1 True ancestral sequences

Generally, the true value of reconstructed ancestral sequences is unknown. The available data are typically sequences from extant (currently existing) species, might not provide a comprehensive representation of the entire evolutionary history. In most cases, the true value of the ancestral species (got extinct) does not exist with few exceptions [Krause-Kyora et al. 2018, Kocher et al. 2021]. Incomplete or sparse data may lead to bias in reconstructing ancestral sequences, while all these limitations lead to inference uncertainty. Alongside all the debates about the current ASR niche, the absence of data with a true ancestral state results in theoretical and methodological limitations. Currently, simulated data under a semi-realistic evolution model is the ultimate solution.

6.2.2 Realistic model of indels

The PIP model is a single-character indel process with constant substitution and indel rate (single rate point indel model). Chapter 4 explained that the single-character PIP model performs reasonably in the ASR task, while ASR under the evolutionary model of multiple-character indels has not been practically demonstrated. Current ASR methods need a more realistic model

of substitution and indel that is capable of capturing rate variation of substitution (e.g., using branch and bound mechanism [Pupko et al. 2002]) and indel (e.g., using GeoPIP model [Zhai and Bouchard-Côté 2017]).

Regarding gap modeling, we need a realistic indel model with an explicit biological model of long indel (inference of long insertion and long deletions). Considering closest extended neighbor sequences [Paten et al. 2008] or adding structural information such as protein secondary structure can improve indel reconstruction in ASR.

6.2.3 Bias of the estimator

The ultimate goal of ASR is to infer the ancestral sequence as close as possible to the true value A. As explained in the introduction, we estimate the set of parameters $\hat{\theta}$ instead of true value θ (θ are parameters such as branch length, tree topology, etc.). We can define the bias of an estimator as the difference between the estimator's expected value $E(\hat{\theta})$ and the true value θ (i.e., $Bias(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$). As we expect, this estimation gives us an error term measured by various loss functions. Notice that there is a trade-off between random errors (captured by variance) and systematic errors (captured by bias). Here, we discuss estimator bias, which is different from the concept of bias in ASR (i.e., selecting the optimal MPA instead of sub-optimal MPAs).

It has been known that the inferred ancestral sequences are biased toward the alignment and tree inference. For example, [Royer-Carenzi et al. 2013] showed that ASR performance relies heavily on tree topology while [Gascuel and Steel 2014] investigated the effect of trees on predicting ancestral changes. Conditioning the ASR on inaccurate MSA or tree will introduce biases to the inference. Different models might yield different results, carrying a level of uncertainty regarding the accuracy of the inferred ancestral sequences. This means the quality of the tree and MSA is crucial for the reliability of ASR.

One possible solution to reduce this bias is the joint estimation of the tree parameters, MSA, and ancestral sequences. Inferring the MSA, tree, and ASR simultaneously makes the estimation less biased. However, the computational aspect of this solution is the bottleneck, as this iterative procedure requires joint inference of many parameters (see 6.3.3 section).

Another possible solution is to reduce this bias by parsimony joint estimation using a state-of-the-art method for each step. This thesis used methods with tuned parameters to provide us with the locally best solution to reduce such a bias in the ASR inference. Using highly tuned statistical approaches (e.g., aligners such as PRANK and tree reconstructors like PhyML) provides us with reasonable results.

6.3 Future research works

This thesis concludes by suggesting research avenues building on top of the current piece of research.

6.3.1 Marginal ASR with indel under PIP

This thesis focused on joint ASR. Yang et al. [1995] explained that one can marginalize over other nodes to compute the ancestral sequence at the node of interest. After a simple adaptation of the DP algorithm, the marginal ASR under PIP is possible. It would be desirable to add it to ARPIP.

6.3.2 Comparative study: joint vs. marginal ASR

It has been reported that joint and marginal reconstruction performance is virtually identical [Gascuel and Steel 2014]. However, there are pieces of evidence that joint and marginal reconstruction results are not always the same [Pupko et al. 2000]. Despite some reports that the results' differences are insignificant, it remains an open research question as no comparative study has been performed. The choice of joint or marginal reconstruction depends on the researcher's intention. For example, joint reconstruction was chosen because we were interested in multiple squamate ancestral nodes in chapter 5. A molecular restoration study is an example of marginal reconstruction. Assuming all the ancestral nodes are subject to the study, it is still unknown which type of reconstruction (joint or full marginal) gives us the best result. To the writer's knowledge, no systematic study demonstrates whether joint or marginal reconstruction is better. One possible research path is to benchmark the two approaches on both real and simulated data.

6.3.3 Simultaneous estimation of tree, MSA, and ancestral sequence with an explicit model of indels

In the current version of ARPIP, we have an efficient ASR method with an explicit evolutionary model of indels. Aligned FASTA-format sequences are the minimum input for the current version of ARPIP. It is possible to integrate an existing MSA tool with/without modeling of indel to ARPIP, so the minimum input for the software would be unaligned sequences. However, this is not the ultimate goal of the ASR task.

If a tree reconstruction tool with an explicit model of indels is developed, the simultaneous estimation of MSA, tree, and ancestral sequences would be possible. In this scenario, the resulting join inference of MSA, tree, and ancestral sequence reconstruction provides us with a locally

optimal solution. Any hill climbing method could explore this optimization space for a globally optimum solution. However, we never know if we will find the global optimum solution. Heuristic techniques and algorithmic methods (such as DP) were deployed to find the MPA sequence in a reasonable time. This way, we reduced the search time for MPA sequence inference to linear time, but still no guarantee of finding the globally optimal solution.

From a computational complexity viewpoint, the problem of ASR is NP-hard. Here, we assumed a known tree topology (and branch lengths) and tried to infer only the ancestral sequences. Current solutions to this problem are heuristics trying to find optimal or approximate solutions [Elias and Tuller 2007].

Joint MSA, tree, and ASR estimation could potentially reduce the inference bias. This is a combinatorial optimization problem, as each MSA, tree, and ASR step needs a separate computation. This computation within the ML framework is possible.

6.3.4 ASR with substitution and indel rate heterogeneity

Heterogeneity in an evolutionary model refers to the fact that the different parts of a biological sequence may evolve at different rates or under different selection mechanisms (substitution or indel). Much evidence exists concerning the substitution rate variation across alignment sites [Cai et al. 2004]. In some organisms (e.g., viruses), the PIP assumption of constant indel rate is justified [Bouchard-Côté and Jordan 2013]. ASR methods need to consider the among-site rate variation as Yang showed that site rate variation can significantly improve the model fit [Yang 1993].

Similar to some evolutionary models, PIP has a biologically unrealistic assumption that different sites in the MSA evolve at the same rate. In reality, this assumption does not hold for indel and substitution since different regions of sequences evolve at different rates due to functional and structural constraints [Yang 2014, p. 15]. To simplify the model (reduce the number of parameters) and make the computation efficient, PIP assumes the process is homogeneous (as well as other CTMC assumptions, reversibility, and stationary).

In chapter 4, we demonstrated that the current single-character version of ARPIP is adequate in the ASR task, while ASR under the evolutionary model of multiple-character indels mainly remains a theoretical question. Similar studies in phylogenetic tree inference provide evidence that one potential factor behind it is that single rate point indel models tend to be detrimental for the inference [Zhai and Bouchard-Côté 2017].

To improve the accuracy of ancestral inference, it is crucial to consider site variability. Without integrated models for indel together with substitutions, the method carries potential biases [Sel-

berg et al. 2021]. The second version of FastML Ashkenazy et al. [2012] used the branch-and-bound algorithm that considered variability of substitution (heterogeneity for the substitution process) when computing the joint reconstruction [Pupko et al. 2002]. However, there are not enough studies on the indel process. One possibility is to extend the current PIP model of indels with indel rate variation, for example, by introducing two or more discrete categories of sites (e.g., "fast" and "slow" indel-rate sites). Sites that are important to protein function are usually highly conserved (slow indel rate sites), while other sites evolve at higher rates (fast indel rate sites) [Pupko et al. 2007]. This implies that indel-enriched sites have different evolution patterns. Therefore, different sites need different substitution and indel rates.

Even though rate variation among sites has not been considered in the current version of ARPIP, one could produce a version with a gamma distribution as proposed in Pupko et al.'s branch-and-bound algorithm [Pupko et al. 2002]. In addition to the mentioned technique, various heuristic methods (such as segmentation) can be used. Pupko et al. [2002] show that among-site variation increases the accuracy of MPA for highly divergent sequences. Regarding computation, the marginal reconstruction assuming site variation is linear but not the same for the joint reconstruction. The current proposed approach for joint ASR assumes site variation is exponential in the worst case with respect to the number of sequences [Pupko et al. 2007].

Traditionally, indel lengths were assumed to obey geometric distribution (the affine gap model [Gotoh 1982]) and, therefore, a mixture of geometric distribution has been proposed to describe the gap length distribution in MSAs [Lunter 2007]. Later, many empirical studies revealed that Zipfian (power-law) distribution explains the data well [Cartwright 2009]. Zhai and Bouchard-Côté [2017] introduced a geometric Poisson indel process (GeoPIP) that allows indel rates to vary across sites. They proposed indel-aware phylogenetic reconstruction methods using GeoPIP. Our proposed solution is to extend this rate variation by branch-and-bound technique to include indel rate variation over time.

Different approaches could be proposed to improve ARPIP without sacrificing its computation efficiency.This could be extended to a generalized ASR method designed for long indels as proposed in [Zhai and Bouchard-Côté 2017]. I believe that extending PIP to GeoPIP could improve the reconstruction performance of ARPIP with respect to the gap/indel length distributions.

6.3.5 Investigating the correlation of indel rates and substitution rates

Evolutionary models assume indels are independent of substitutions, but this may not be true. Many researchers studied the dynamic of substitutions, such as positive selection (e.g., [Vallender and Lahn 2004, Arbiza et al. 2006, Kosiol et al. 2008]). In chapter 4, different metrics have been proposed to investigate the indel dynamic in mammalian proteins. Correlation of indel and

substitution rate is another interesting future direction to explore [Jovelin and Cutter 2014, Zhai 2016, Gonnet et al. 1992]. The possible outcome would result in a new brand of evolutionary models considering the behavior of substitution and indels simultaneously.

6.3.6 Using structural and physicochemical-based information

Reconstructed ancestral sequences should be stable, foldable, and functional [Pupko et al. 2007]. The ultimate goal of ASR is to have a highly accurate MPA that satisfies this biological requirement. Studies used structural and physicochemical evolutionary models to reconstruct [Taverna and Goldstein 2002] and study the accuracy of reconstruction [Williams et al. 2006]. This information can be used to improve the existing evolutionary models and increase inference accuracy to develop a more realistic ASR.

6.3.7 Inferring the evolutionary history of sequences with deep neural networks

With advances in sequencing technologies, vast amounts of molecular information are produced, contributing to the big data era in biology. Many computational tools rely on machine learning and extensive databases for training data. Machine learning techniques, such as neural networks (NN), can capture more complex relationships (non-linear patterns) in the data. In most cases, the quality of the data available determines the accuracy of such tools [Thomson et al. 2022]. However, it is challenging to propose a machine-learning tool to infer the ancestral sequence.

The lack of true real-world ancestral data is a limitation of having a supervised machine-learning method. One possibility is to train these models on simulated data from statistical simulator packages such as INDELible [Fletcher and Yang 2009] and JavaPIP. The model parameters of such simulations could be tuned so that the generated sequences resemble real data (as has been done in chapter 3 of this thesis). As simulation data are increasingly used to validate ASR method [Vialle et al. 2018, Jowkar et al. 2023, Foley et al. 2022], we can use this type of data to train a neural network.

The lack of true ancestral sequence is a big challenge for ASR using most of machine learning methods, which rely heavily on big labeled data. In most cases, we only have present-day sequences with few exceptions (e.g., [Krause-Kyora et al. 2018, Kocher et al. 2021]). This thesis suggests developing unsupervised or semi-supervised machine learning methods instead. In the context of neural networks, each node of the tree can be considered as the neurons of our network. Edges between two nodes in the tree represent the time difference or the amount of change between them in the form of a distance matrix (for e.g., patristic distance matrix

[Moreta et al. 2021]). A training objective could be to minimize the negative log-likelihood of the observed data given the model. With existing compatible hardware and infrastructure, this is possible.

Bibliography

Altenhoff, A. M. and Dessimoz, C. (2012). Inferring orthology and paralogy. *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, pages 259–279.

Altenhoff, A. M., Glover, N. M., and Dessimoz, C. (2019a). Inferring orthology and paralogy. *Evolutionary genomics: statistical and computational methods*, pages 149–175.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., et al. (2019b). Oma standalone: orthology inference among public and custom genomes and transcriptomes. *Genome research*, 29(7):1152–1163.

Altenhoff, A. M., Train, C.-M., Gilbert, K. J., Mediratta, I., Mendes de Farias, T., Moi, D., Nevers, Y., Radoykova, H.-S., Rossier, V., Warwick Vesztrocy, A., et al. (2021). Oma orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic acids research*, 49(D1):D373–D379.

Arbiza, L., Dopazo, J., and Dopazo, H. (2006). Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS computational biology*, 2(4):e38.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., and Pupko, T. (2012). Fastml: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research*, 40(W1):W580–W584.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., et al. (2020). Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D480–D489.

Belouzard, S., Millet, J. K., Licitra, B. N., and Whittaker, G. R. (2012). Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses*, 4(6):1011–1033.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2):163–193.

Bokeh Development Team (2018). *Bokeh: Python library for interactive visualization*.

Bouchard-Côté, A. (2010). *Probabilistic Models of Evolution and Language Change*. PhD thesis, University of California at Berkeley, University of California at Berkeley. PhD thesis.

Bouchard-Côté, A. and Jordan, M. I. (2013). Evolutionary inference via the poisson indel process. *Proceedings of the National Academy of Sciences*, 110(4):1160–1166.

Brent, R. P. (1973). Algorithms for minimization without derivatives. *Englewood Cliffs, NJ, USA Prentice Hall*, page 195.

Brintnell, E., Gupta, M., and Anderson, D. W. (2021). Phylogenetic and ancestral sequence reconstruction of sars-cov-2 reveals latent capacity to bind human ace2 receptor. *Journal of molecular evolution*, 89(9):656–664.

Britten, R. J., Rowen, L., Williams, J., and Cameron, R. A. (2003). Majority of divergence between closely related dna samples is due to indels. *Proceedings of the National Academy of Sciences*, 100(8):4661–4665.

Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. (1982). Mitochondrial dna sequences of primates: tempo and mode of evolution. *Journal of molecular evolution*, 18:225–239.

Cai, W., Pei, J., and Grishin, N. V. (2004). Reconstruction of ancestral protein sequences and its applications. *BMC evolutionary biology*, 4:1–23.

Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular biology and evolution*, 26(2):473–480.

Čerňanskỳ, A., Boistel, R., Fernandez, V., Tafforeau, P., Nicolas, L. N., and Herrel, A. (2014). The atlas-axis complex in chamaeleonids (squamata: Chamaeleonidae), with description of a new anatomical structure of the skull. *The Anatomical Record*, 297(3):369–396.

Chang, B. S., Ugalde, J. A., and Matz, M. V. (2005). Applications of ancestral protein reconstruction in understanding protein function: Gfp-like proteins. In *Methods in enzymology*, volume 395, pages 652–670. Elsevier.

Chuzhanova, N. A., Anassis, E. J., Ball, E. V., Krawczak, M., and Cooper, D. N. (2003). Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local dna sequence complexity. *Human mutation*, 21(1):28–44.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). *Introduction to algorithms*. MIT press.

Cunningham, C. W., Zhu, H., and Hillis, D. (1998). Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, 52(4):978–987.

de Jong, W. W. and Rydén, L. (1981). Causes of more frequent deletions than insertions in mutations and protein evolution. *Nature*, 290(5802):157–159.

De Maio, N. (2021). The cumulative indel model: fast and accurate statistical evolutionary alignment. *Systematic Biology*, 70(2):236–257.

Dessimoz, C. and Gil, M. (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome biology*, 11(4):R37.

Diallo, A. B., Makarenkov, V., and Blanchette, M. (2007). Exact and heuristic algorithms for the indel maximum likelihood problem. *Journal of Computational Biology*, 14(4):446–461.

Diallo, A. B., Makarenkov, V., and Blanchette, M. (2009). Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1):130–131.

Edwards, R. J. and Shields, D. C. (2004). Gasp: gapped ancestral sequence prediction for proteins. *BMC bioinformatics*, 5:1–10.

Elias, I. and Tuller, T. (2007). Reconstruction of ancestral genomic sequences using likelihood. *Journal of Computational Biology*, 14(2):216–237.

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

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416.

Fletcher, W. and Yang, Z. (2009). Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888.

Foley, G., Mora, A., Ross, C. M., Bottoms, S., Sützl, L., Lamprecht, M. L., Zaugg, J., Essebier, A., Balderson, B., Newell, R., et al. (2022). Engineering indel and substitution variants of diverse and ancient enzymes using graphical representation of ancestral sequence predictions (grasp). *PLOS Computational Biology*, 18(10):e1010633.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., et al. (2002). Diversity considerations in hiv-1 vaccine selection. *Science*, 296(5577):2354–2360.

Gascuel, O. and Steel, M. (2014). Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Systematic biology*, 63(3):421–435.

Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2008). Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature*, 451(7179):704–707.

Gauthier, J. A., Kearney, M., Maisano, J. A., Rieppel, O., and Behlke, A. D. (2012). Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. *Bulletin of the Peabody Museum of Natural History*, 53(1):3–308.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256(5062):1443–1445.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3):705–708.

Groussin, M., Hobbs, J. K., Szöllősi, G. J., Gribaldo, S., Arcus, V. L., and Gouy, M. (2014). Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees. *Molecular Biology and Evolution*, 32(1):13–22.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., et al. (2013). Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular biology and evolution*, 30(8):1745–1750.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic biology*, 59(3):307–321.

Hao, H., Tummala, P., Guzman, E., Mali, R. S., Gregorski, J., Swaroop, A., and Mitton, K. P. (2011). The transcription factor neural retina leucine zipper (nrl) controls photoreceptor-specific expression of myocyte enhancer factor mef2c from an alternative promoter. *Journal of Biological Chemistry*, 286(40):34893–34902.

He, Y., Tian, S., and Tian, P. (2019). Fundamental asymmetry of insertions and deletions in genomes size evolution. *Journal of Theoretical Biology*, 482:109983.

Holmes, I. (2020). A model of indel evolution by finite-state, continuous-time machines. *Genetics*, 216(4):1187–1204.

Holmes, I. H. (2017). Historian: accurate reconstruction of ancestral sequences and evolutionary rates. *Bioinformatics*, 33(8):1227–1229.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638.

Jensen, R. A. (2001). Orthologs and paralogs-we need to get it right. *Genome biology*, 2(8):interactions1002–1.

Jovelin, R. and Cutter, A. D. (2014). Microevolution of nematode mirnas reveals diverse modes of selection. *Genome biology and evolution*, 6(11):3049–3063.

Jowkar, G., Pečerska, J., Maiolo, M., Gil, M., and Anisimova, M. (2023). Arpip: Ancestral sequence reconstruction with insertions and deletions under the poisson indel process. *Systematic biology*, 72(2):307–318.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., and Poon, A. F. (2016). Ancestral reconstruction. *PLoS computational biology*, 12(7):e1004763.

Kamneva, O. K., Liberles, D. A., and Ward, N. L. (2010). Genome-wide influence of indel substitutions on evolution of bacteria of the pvc superphylum, revealed using a novel computational method. *Genome Biology and Evolution*, 2:870–886.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., et al. (2014). Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31:151–160.

Kocher, A., Papac, L., Barquera, R., Key, F. M., Spyrou, M. A., Hübler, R., Rohrlach, A. B., Aron, F., Stahl, R., Wissgott, A., et al. (2021). Ten millennia of hepatitis b virus evolution. *Science*, 374(6564):182–188.

Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of molecular evolution*, 42(2):313–320.

Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six mammalian genomes. *PLoS genetics*, 4(8):e1000144.

Krause-Kyora, B., Susat, J., Key, F. M., Kühnert, D., Bosse, E., Immel, A., Rinne, C., Kornell, S.-C., Yepes, D., Franzenburg, S., et al. (2018). Neolithic and medieval virus genomes reveal complex evolution of hepatitis b. *Elife*, 7:e36666.

Kuo, C.-H. and Ochman, H. (2009). Deletional bias across the three domains of life. *Genome biology and evolution*, 1:145–152.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320.

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., and Smith, D. B. (2018). Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, 46(D1):D708–D717.

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, 34(9):2422–2424.

Liberles, D. A. (2007). *Ancestral sequence reconstruction*. Oxford University Press on Demand.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., and Guo, J.-t. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific reports*, 7(1):9313.

Loewenthal, G., Rapoport, D., Avram, O., Moshe, A., Wygoda, E., Itzkovitch, A., Israeli, O., Azouri, D., Cartwright, R. A., Mayrose, I., et al. (2021). A probabilistic model for indel evolution: differentiating insertions from deletions. *Molecular biology and evolution*, 38(12):5769–5781.

Löytynoja, A. (2014). Phylogeny-aware alignment with prank. In *Multiple sequence alignment methods*, pages 155–170. Springer.

Löytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences*, 102(30):10557–10562.

Lunter, G. (2007). Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13):i289–i296.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2003). Bayesian phylogenetic inference under a statistical insertion-deletion model. In *Algorithms in Bioinformatics: Third International Workshop, WABI 2003, Budapest, Hungary, September 15-20, 2003. Proceedings 3*, pages 228–244. Springer.

Macdonald, C. B., Nedrud, D., Grimes, P. R., Trinidad, D., Fraser, J. S., and Coyote-Maestas, W. (2022). Deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. *bioRxiv*.

Maiolo, M. (2019). *Progressive Multiple Sequence Alignment with Indel Evolution*. PhD thesis, University of Lausanne, Lausanne. PhD thesis.

Maiolo, M., Gatti, L., Frei, D., Leidi, T., Gil, M., and Anisimova, M. (2021). Propip: a tool for progressive multiple sequence alignment with poisson indel process. *BMC bioinformatics*, 22:1–12.

Maiolo, M., Zhang, X., Gil, M., and Anisimova, M. (2018). Progressive multiple sequence alignment with indel evolution. *BMC bioinformatics*, 19(1):331.

McKee, S. P., McCann, J. J., and Benton, J. L. (1977). Color vision from rod and long-wave cone interactions: Conditions in which rods contribute to multicolored images. *Vision research*, 17(2):175–185.

Miklós, I., Lunter, G. A., and Holmes, I. (2004). A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–540.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome research*, 16(9):1182–1190.

Minuchehr, Z. and Goliaei, B. (2005). Propensity of amino acids in loop regions connecting beta-strands. *Protein and peptide letters*, 12(4):379–382.

Montana, C. L., Lawrence, K. A., Williams, N. L., Tran, N. M., Peng, G.-H., Chen, S., and Corbo, J. C. (2011). Transcriptional regulation of neural retina leucine zipper (nrl), a photoreceptor cell fate determinant. *Journal of Biological Chemistry*, 286(42):36921–36931.

Moreta, L. S., Rønning, O., Al-Sibahi, A. S., Hein, J., Theobald, D., and Hamelryck, T. (2021). Ancestral protein sequence reconstruction using a tree-structured ornstein-uhlenbeck variational autoencoder. In *International Conference on Learning Representations*.

Nee, S., Holmes, E. C., May, R. M., and Harvey, P. H. (1994). Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1307):77–82.

Nevers, Y., Glover, N. M., Dessimoz, C., and Lecompte, O. (2023). Protein length distribution is remarkably uniform across the tree of life. *Genome Biology*, 24(1):135.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7):358–364.

Ogurtsov, A. Y., Sunyaev, S., and Kondrashov, A. S. (2004). Indel-based evolutionary distance and mouse–human divergence. *Genome Research*, 14(8):1610–1616.

Oliva, A., Pulicani, S., Lefort, V., Brehelin, L., Gascuel, O., and Guindon, S. (2019). Accounting for ambiguity in ancestral sequence reconstruction. *Bioinformatics*, 35(21):4290–4297.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., and Birney, E. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research*, 18(11):1829–1843.

Pečerska, J., Gil, M., and Anisimova, M. (2021). Joint alignment and tree inference. *bioRxiv*.

Pollock, D. D. and Chang, B. S. (2007). Dealing with uncertainty in ancestral sequence reconstruction: sampling from the posterior distribution. *Ancestral sequence reconstruction*, 85:94.

Pupko, T., Doron-Faigenboim, A., Liberles, D. A., and Cannarozzi, G. M. (2007). *Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences*, volume 4. chapter.

Pupko, T., Pe, I., Shamir, R., and Graur, D. (2000). A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., and Friedman, N. (2002). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics*, 18(8):1116–1123.

Redelings, B. D. (2021). Bali-phy version 3: model-based co-estimation of alignment and phylogeny. *Bioinformatics*, 37(18):3032–3034.

Ribeiro-dos Santos, A. M., da Silva, V. L., de Souza, J. E., and de Souza, S. J. (2015). Populational landscape of indels affecting transcription factor-binding sites in humans. *BMC genomics*, 16:1–11.

Ross, C. M., Foley, G., Boden, M., and Gillam, E. M. (2022). Using the evolutionary history of proteins to engineer insertion-deletion mutants from robust, ancestral templates using graphical representation of ancestral sequence predictions (grasp). *Enzyme engineering: methods and protocols*, pages 85–110.

Ross, H. A., Nickle, D. C., Liu, Y., Heath, L., Jensen, M. A., Rodrigo, A. G., and Mullins, J. I. (2006). Sources of variation in ancestral sequence reconstruction for hiv-1 envelope genes. *Evolutionary Bioinformatics*, 2:117693430600200027.

Royer-Carenzi, M., Pontarotti, P., and Didier, G. (2013). Choosing the best ancestral character state reconstruction method. *Mathematical Biosciences*, 242(1):95–109.

Ruse, M. (1975). Charles darwin's theory of evolution: an analysis. *Journal of the History of Biology*, pages 219–241.

Schott, R. K., Müller, J., Yang, C. G., Bhattacharyya, N., Chan, N., Xu, M., Morrow, J. M., Ghenu, A.-H., Loew, E. R., Tropepe, V., et al. (2016). Evolutionary transformation of rod photoreceptors in the all-cone retina of a diurnal garter snake. *Proceedings of the National Academy of Sciences*, 113(2):356–361.

Sehn, J. K. (2015). Insertions and deletions (indels). In *Clinical genomics*, pages 129–150. Elsevier.

Selberg, A. G., Gaucher, E. A., and Liberles, D. A. (2021). Ancestral sequence reconstruction: from chemical paleogenetics to maximum likelihood algorithms and beyond. *Journal of Molecular Evolution*, 89:157–164.

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

Simões, B. F., Sampaio, F. L., Douglas, R. H., Kodandaramaiah, U., Casewell, N. R., Harrison, R. A., Hart, N. S., Partridge, J. C., Hunt, D. M., and Gower, D. J. (2016). Visual pigments, ocular filters and the evolution of snake vision. *Molecular biology and evolution*, 33(10):2483–2495.

Simoes, B. F., Sampaio, F. L., Loew, E. R., Sanders, K. L., Fisher, R. N., Hart, N. S., Hunt, D. M., Partridge, J. C., and Gower, D. J. (2016). Multiple rod–cone and cone–rod photoreceptor transmutations in snakes: evidence from visual opsin gene expression. *Proceedings of the Royal Society B: Biological Sciences*, 283(1823):20152624.

Söding, J. and Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, 25(9):837–846.

Spence, M. A., Kaczmarski, J. A., Saunders, J. W., and Jackson, C. J. (2021). Ancestral sequence reconstruction for protein engineers. *Current Opinion in Structural Biology*, 69:131–141.

Starr, T. N., Zepeda, S. K., Walls, A. C., Greaney, A. J., Alkhovsky, S., Veesler, D., and Bloom, J. D. (2022). Ace2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature*, pages 1–9.

Swaroop, A., Kim, D., and Forrest, D. (2010). Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nature Reviews Neuroscience*, 11(8):563–576.

Tao, S., Fan, Y., Wang, W., Ma, G., Liang, L., and Shi, Q. (2007). Patterns of insertion and deletion in mammalian genomes. *Current Genomics*, 8(6):370–378.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of dna sequence. *Lecture of Mathematics for Life Science*, 17:57.

Taverna, D. M. and Goldstein, R. A. (2002). Why are proteins so robust to site mutations? *Journal of molecular biology*, 315(3):479–484.

Taylor, M. S., Ponting, C. P., and Copley, R. R. (2004). Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome research*, 14(4):555–566.

Thomson, R. E., Carrera-Pacheco, S. E., and Gillam, E. M. (2022). Engineering functional thermostable proteins using ancestral sequence reconstruction. *Journal of Biological Chemistry*, page 102435.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, 33(2):114–124.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of molecular evolution*, 34(1):3–16.

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics*, 5(5):366–375.

Tóth-Petróczy, A. and Tawfik, D. S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. *Molecular biology and evolution*, 30(4):761–771.

Vallender, E. J. and Lahn, B. T. (2004). Positive selection on the human genome. *Human Molecular Genetics*, 13(suppl_2):R245–R254.

Vialle, R. A., Tamuri, A. U., and Goldman, N. (2018). Alignment modulates ancestral sequence reconstruction accuracy. *Molecular biology and evolution*, 35(7):1783–1797.

Walls, G. (1934). The reptilian retina: I. a new concept of visual-cell evolution. *American journal of ophthalmology*, 17(10):892–915.

Wang, S. and Cepko, C. L. (2016). Photoreceptor fate determination in the vertebrate retina. *Investigative ophthalmology & visual science*, 57(5):ORSFe1–ORSFe6.

Westesson, O., Barquist, L., and Holmes, I. (2012). Handalign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics*, 28(8):1170–1171.

Wetterbom, A., Sevov, M., Cavelier, L., and Bergström, T. F. (2006). Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *Journal of molecular evolution*, 63:682–690.

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*, 18(5):691–699.

Williams, P. D., Pollock, D. D., Blackburne, B. P., and Goldstein, R. A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLoS computational biology*, 2(6):e69.

Wygoda, E., Loewenthal, G., Moshe, A., Alburquerque, M., Mayrose, I., and Pupko, T. (2024). Statistical framework to determine indel-length distribution. *Bioinformatics*, 40(2):btae043.

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*, 13(5):197–204.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396–1401.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., et al. (2020). Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688.

Zakas, P. M., Brown, H. C., Knight, K., Meeks, S. L., Spencer, H. T., Gaucher, E. A., and Doering, C. B. (2017). Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nature biotechnology*, 35(1):35.

Zhai, Y. (2016). *Stochastic Processes, Statistical Inference and Efficient Algorithms for Phylogenetic Inference*. PhD thesis, University of British Columbia, University of British Columbia. PhD thesis.

Zhai, Y. and Bouchard-Côté, A. (2017). A poissonian model of indel rate variation for phylogenetic tree inference. *Systematic Biology*, 66(5):698–714.

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

Zhou, G. and Zhao, Q. (2020). Perspectives on therapeutic neutralizing antibodies against the novel coronavirus sars-cov-2. *International journal of biological sciences*, 16(10):1718.

Curriculum Vitae

Gholam-Hossein Jowkar

Born on November 20th, 1989

LinkedIn, GitHub, Google

EDUCATION

2019-2024 **Ph.D. In Computational Biology,** *University of Neuchatel, Switzerland.*

Major: Computational molecular evolutionary biology
Thesis: Frequentist Estimation of the Evolutionary History of Sequences with Substitutions & Indels
Description: The evolutionary history of molecules is described by a tree structure called a phylogeny, inferred from genetic sequences. Phylogenies are used for testing biological hypotheses with applications ranging from medicine to ecologyIn this project I have developed a new maximum likelihood method for reconstructing ancestral sequences on a phylogeny – ARPIP. This method uses a sequence evolution model that accounts for insertions and deletions (indels) and is therefore able to infer more biologically meaningful ancestral sequences. In the later stage of my PhD I have worked on a large-scale indel study of mammalian genomes and also developed an indel visualization.
Two software packages: ARPIP (C++) and IndelViewer (Python)

2013-2016 **M. Sc. In Computer Engineering,** *Shiraz University, Shiraz, Iran*.

Major: Artificial Intelligence and Robotics **with GPA** 16.86/$_{20.00}$
Thesis: Developing a Hybrid Method for Disease Gene Identification
Description: In this project I developed a perceptron ensemble of graph-based positive-unlabeled learning (PEGPUL) method for identifying disease genes. The method used three biological attributes and a positive unlabeled co-learning schema to classify ~12000 disease genes. This project highlights the importance of using machine learning tools for identifying disease genes and demonstrates the potential of semi-supervised learning algorithms on a large dataset of disease genes.

2008-2013 **B. Sc. In Computer Engineering,** *Azad University of Shiraz, Shiraz, Iran*.

Major: Software Engineering **with GPA** 16.49/$_{20.00}$
Final Project: Data mining using Image Processing and Artificial Neural Networks
Description: In this project, I explored the use of morphological features and artificial neural networks to classify four Iranian wheat cultivars. I extracted ten morphological features from 164-grain images for each cultivar and then developed a multilayer perceptron neural network for classification. This study was one of a kind.

2003-2008 **High School**, *Taha High School,* Shiraz, Iran.
Major: Mathematics & Physics

TECHNICAL SKILLS

Programming Languages

Python(proficient): Scikit-learn, Pandas, Numpy, Seaborn, Bokeh, Panel, Tensorflow, PyMongo, BioPython, and Django.

C++(proficient): Bio++, Standard Library, Boost, GLog, GoogleTest, and Design Pattern.

	MATLAB(proficient): Bioinformatics, Image Processing, Signal Processing, Computer Vision, Statistical and Machine Learning, and Text Analytics toolboxes.
	Java(knowledge): Android Studio.
	C# (knowledge): ASP.Net Webform and MVC.
Databases	SQL databases (MS SQL, MySQL, and SQLite), NoSQL (MongoDB: CRUD).
Operating Systems	Linux (proficient in Bash script and working with HPC) and Windows.
Project management	Agile Principles, and Kanban.
Machine learning	Supervised, Unsupervised, Semi-supervised Learning, Dimension Reduction, Feature Selection, Neural networks (Classical and Deep Learning), and Metric Learning.
Other skills	Software System Analysis with UML, LaTeX, R scripting, Linear Optimization, Heuristic and Metaheuristic Optimization Methods (E.g., Genetic Algorithms).
Selected coursework	Machine Learning, Statistical Pattern Recognition, Evolutionary Computation, Fuzzy Logic, Digital Image Processing, Neural Networks, Bioinformatics, Computational Evolutionary Biology, Advanced Mathematics-Linear Algebra, Non-Linear Optimization and Advanced Statistical Pattern Recognition.

PROFESSIONAL EXPERIENCES

2019-present **Research assistant at Zurich University of applied sciences**

- https://www.zhaw.ch/en/about-us/person/jowk/
- Data scientist in Prof. Dr. Anisimova's group mainly focusing on software development and research in computational evolutionary biology.

2018-2019 **Freelancing and working in a startup company**

- I led a group of 3 software developers, focusing on Android mobile applications, web platforms, and AI development. During this period, we have worked on two projects focusing on knowledge extraction from Persian texts and false information detection application using gamification methods.

2017-2019 **Co-founder of Al Khwarizmi Center for Entrepreneurship (NGO)**

- In this NGO, we have helped high school students to prototype their innovative ideas. The aim was to develop minimal valuable products and write a simple business plan. The project had ~50 active members and a crowdfunding-like website for the projects.

2016-2017 **Co-founder and CEO of "Fardan Haft Eghlim" Company**

- www.fardan7eghlim.ir
- Co-Founding a company and managing a team of 4 software developers working on cross-platform software development. In this company, we developed Android applications and websites for private customers.