ExpDist: An Estimator for Expected Values of Evolutionary Distances

Julia Radenholt

Bachelor's Thesis in Computer Science
at Stockholm University, Sweden, 2021

ExpDist: An Estimator for Expected Values of Evolutionary Distances

Julia Radenholt

Bachelor’s Thesis in Computer Science (15 ECTS credits)
Bachelor’s Programme in Computer Science
Stockholm University year 2021
Supervisor at the Department of Mathematics was Lars Arvestad
Examiner was Woosok Moon

Department of Mathematics
Stockholm University
SE-106 91 Stockholm, Sweden

Abstract

Evolutionary history can be reconstructed by protein phylogeny, from data retrieved from sequences of amino acids. The most common methods, maximum likelihood and maximum a posteriori, are using the transition probabilities from an ancestral sequence to a descendant sequence to estimate the evolutionary distance between the sequences. There exists weakness in these methods. For example there exist sequences of amino acids that are identical and are retrieved from different species, but the methods will always estimate the evolutionary distance between identical sequences to zero.

This study aims to determine how it is possible to improve the estimates of the evolutionary distances. Specifically, it investigates whether the discretization in the estimator of the expected values of evolutionary distances can be improved.

The discretization method developed in the study shows it is possible to speed up the computation of the expected values. This finding implies it is possible to implement an efficient estimator of the expected value, which can be used to robust the estimates of evolutionary distances. We therefore introduce a new procedure, called ExpDist, for fast and accurate estimates for the expected values of evolutionary distances between sequences of amino acids.

ExpDist: En uppskattare för väntevärden av evolutionära avstånd

Sammanfattning

Evolutionär historia kan spåras med hjälp av protein fylogeni. Protein fylogeni kan rekonstrueras genom att uppskatta evolutionära avstånd mellan sekvenser av aminosyror. För att uppskatta evolutionära avstånd används vanligtvis maximum likelihood-metoden och a posteriori-fördelningar. Metoderna uppskattar ett evolutionärt avstånd mellan sekvenser genom att parvis linjera sekvenserna och beräkna överggångssannolikheterna mellan aminosyrorna i sekvenserna. Det existerar svagheter i dessa metoder, till exempel är det möjligt att identiska sekvenser har hämtas från olika arter, trots det kommer dessa metoder att uppskatta det evolutionära avståndet mellan sådana sekvenser till noll.

Syftet med denna studie är att undersöka hur det är möjligt att förbättra uppskattningar av evolutionära avstånd, specifikt undersöker vi om diskretiseringen i uppskattaren av väntevärden för evolutionära avstånd kan förbättras.

Diskretiseringsmetoden som utvecklats i studien visar att beräkningar av väntevärden kan göras mer effektivt. Detta resultat antyder att det är möjligt att implementera en effektiv uppskattare av väntevärden för evolutionära avstånd, vilket kan användas för att förbättra de evolutionära avstånd som uppskattas med maximum likelihood eller andra metoder. Vi introducerar därför en ny metod, ExpDist, för snabba och riktiga uppskattningar av väntevärden för evolutionära avstånd mellan sekvenser av aminosyror.

Acknowledgements

I would like to thank my supervisor, Lars Arvestad, for his patience and for providing guidance, feedback and the right phylogenetic tools throughout this project.

Contents

1 Introduction 7

2 Theoretical framework 8

2.1 Sequence alignments and phylogenetic trees 8

2.2 Substitution models 10

2.3 Probability theory 12

2.3.1 Definitions and theorems 12

2.3.2 Probabilites in biological evolution 14

2.3.3 Probability methods 15

2.3.4 Bayesian inference 16

3 Problem 19

3.1 Motivation 19

3.1.1 ExpDist 19

3.2 Related work 19

4 Method 20

4.1 Generating sequence data 20

4.2 Methods for numerical integration 21

4.3 Probabilites 22

4.4 Discretization 22

4.4.1 Test cases 23

4.4.2 Discretization method 24

5 Result and analysis 25

5.1 Limitations in the implementation 29

6 Conclusions 30

6.1 Suggestion for further research 30

References 31

1 Introduction

This thesis explores how it is possible to improve the reconstruction of biological evolution using the conditional expected values of evolutionary distances. We will look into the basics of protein phylogeny, probability in biological evolution and some probability methods. We will study the problem which is that for the most common probability methods there exist weakness in the prediction of the evolution. Two of these methods are maximum likelihood and maximum a posteriori. We will discuss why the expected value of evolutionary distances are necessary and how it is possible to implement an efficient procedure for the expected value. The thesis introduces ExpDist, an efficient estimator for the expected value of evolutionary distances between sequences of amino acids.

2 Theoretical framework

This chapter contains information about the basics of protein phylogeny, probability in biological evolution and probability methods.

2.1 Sequence alignments and phylogenetic trees

An approximation of the evolutionary distance between two sequences of amino acids is regularly estimated by aligning the sequences, observing the number of sites the sequences differ and assigning a cost for each mutation. The cost of a mutation between a pair of amino acids usually depends on which kind of amino acids the mutation occur between. Example of a sequence alignment can be seen at Figure 1.

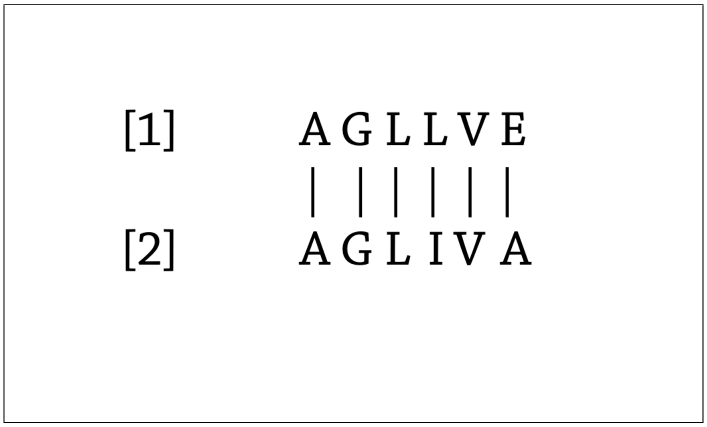


Figure 1: *A sequence alignment. The number of sites that differ is two and all of the pairwise mutations $A \to A$, $G \to G$, $L \to L$, $L \to I$, $V \to V$ and $E \to A$ are assigned with a cost in the estimation of the evolutionary distance from sequence [1] to sequence [2].*

When sequences of amino acids are aligned and the length of the sequences are not equal, there exist gaps in the alignment, see Figure 2. Usually, the gaps are replaced by amino acids or the amino acids in the longer sequences are deleted (Garrett and Grisham 2014).

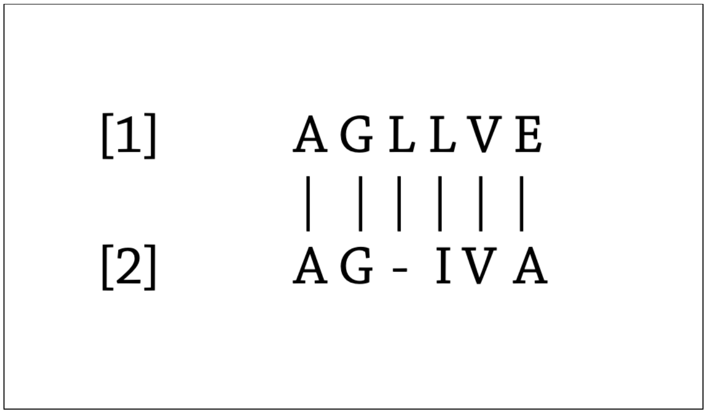


Figure 2: *A sequence alignment with a gap.*

Visualization of evolutionary history can be reconstructed with a phylogenetic tree. A phylogenetic tree is a diagram that depicts the branching history between species, organisms or genes of common ancestry (Baum 2008). A phylogenetic tree can be seen in Figure 3.

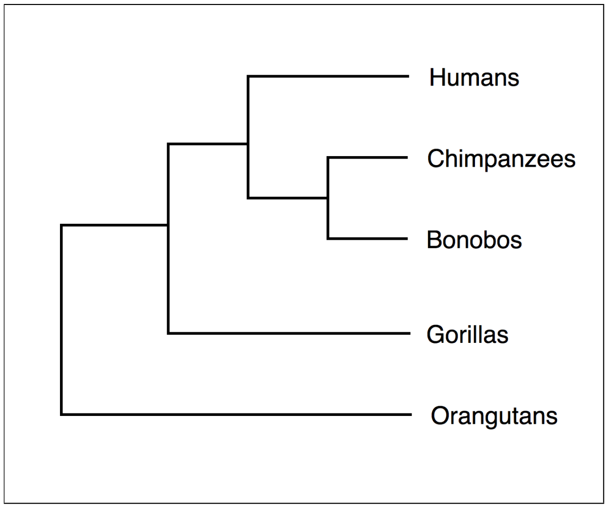


Figure 3: *A phylogenetic tree.*

2.2 Substitution models

At the simplest level, the proportion of sites p where the amino acids have not been conserved can be used to measure the evolutionary distance between two sequences. This proportion is called the p-distance and can be measured by:

$$\hat{p} = \frac{n_d}{n},$$

where n is the total number of amino-acids in the sequence and n_d is the number of different amino acid for the pair.

The Poisson correction distance is another measurement for evolutionary distance at a simpler lever. It assumes the probability of mutation among the sites follows a Poisson distribution, with an uniform rate per site per time unit. The Poisson correction distance can be estimated by a formula which takes the p-distance $\hat{p}$ as input (Nei and Zhang 2005). Poisson correction distance can be measured by:

$$p = -\ln{(1 - \hat{p})}.$$

If p is small, the p-distance approximately is equal to the number of substitutions per site. If p is large, there may be multiple substitutions at a given site, so the p-distance will give an underestimate of the number of substitutions. There exists a number of correction methods, based on probabilistic models, which have been developed to give a more accurate estimate for the p-distance and the Poisson correction distance.

Some of these models are PAM (Dayhoff, Schwartz, and Orcutt 1967), JTT (Jones, Taylor, and Thornton 1992), WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008).

The process of substitution in the models are described with continuous-time Markov chains using matrices of substitution rate and vectors of equilibrium frequencies.

The relationship between sequence divergence and the Poisson correction distance can be seen in Figure 4.

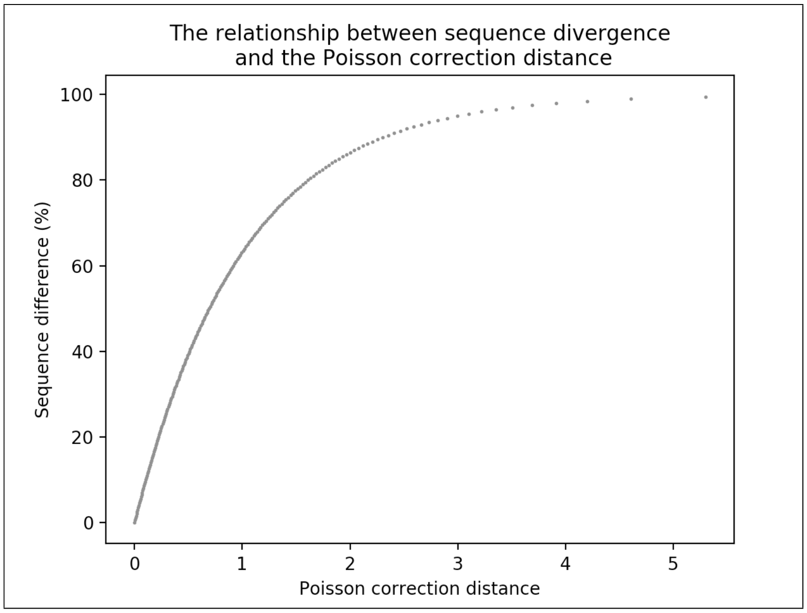


Figure 4: *The relationship between sequence divergence and the Poisson correction distance. The plot shows a rough estimate of the evolutionary distance in relation to the difference in the sequences. The plot also show the existence of the uncertainty in Poisson correction distances for greater values.*

2.3 Probability theory

In this section, definitions and theorems are presented. The definitions and theorems are necessary to understand how it is possible to compute the expected value of an evolutionary distance given a sequence alignment.

2.3.1 Definitions and theorems

In the theory of probabilistic, a random variable, also known as stochastic variable, is a variable for which the values depend on the outcome of a random experiment. The random variable can be discrete or continuous. A discrete random variable takes a finite set of discrete values. A continuous random variable takes on values that vary continuously within one or more real intervals. The set of possible values in a random experiment is called the sample space. Certain subsets of the sample space of an experiment are referred to as events.

The notation $P(X)$ refers to the probability that event X occurs and $P(X|Y)$ refers to the likelihood that event X occurs, given that event Y occurred.

Each of the following definitions are presented both for discrete and continuous random variables. This is due to the fact that this study relies on research where the expected values are found from discrete random variables and in the study are implemented as continuous.

Let us look at the definitions of the expected value.

Definition 2.1. (Expected value)
Suppose X is a discrete random variable that takes values $x_1, x_2, ..., x_n$ with probabilities $P(x_1), P(x_2), ..., P(x_n)$. The expected value of X is denoted $E(X)$ and is given by

$$E(X) = \sum_{j=0}^{n} P(x_j)x_j = P(x_1)x_1 + P(x_2)x_2 + ... + P(x_n)x_n.$$

If X is a continuous random variable with probability density function $f(x)$, then the expected value $E(X)$ is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

The conditional expectation of X given that Y=y is the weighted average of the values that X can take, where each possible is weighted by its respective conditional probability that $Y = y$. The expectation of a discrete random variable X conditional on Y=y is denoted by

$$E(X|Y = y) = \sum_{j=0}^{n} P(x_j|Y = y)x_j,$$

The expectation of a continuous random variable X conditional on Y=y is denoted by

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx.$$

We observe that to estimate the conditional expectation, the posterior distribution has to be known. To compute the posterior distribution, we can use Bayes theorem.

Theorem 1. (Bayes' Theorem)
For a sample space Ω consisting of disjoint events X, with probability $P(X_i) > 0$ for $i = 1, .., n$, such that $\cup_{i=1}^{n} X_i = \Omega$, the probability for any event $Y = \cup_{i=1}^{n}(Y \cap X_i)$ occurring is given by

$$P(X|Y) = \frac{P(X) \cdot P(Y|X)}{P(Y)}.$$

The distribution of the prior probability of the evolutionary distances is assumed to be uniform and the likelihood of the alignment given an evolutionary distance is possible to retrieve from substitution models (See details in section 2.3.1). To be able to use Bayes' theorem, we have to find the prior distribution of the alignments. Since the likelihood of the alignment and the prior probability of the distances are known, we can use the law of total probability to retrieve the probability of an alignment.

Theorem 2. (Law of total probability - Discrete random variable)
For a sample space Ω consisting of disjoint events Y_i with probability $P(Y_i) > 0$ for $i = 1, .., n$ such that $\cup_{i=1}^{n} Y_i = \Omega$, the probability of any event $X = \cup_{i=1}^{n}(X \cap Y_i)$ occuring is given by

$$P(X) = \sum_{i=1}^{n} P(X|Y_i)P(Y_i)$$

Theorem 3. (Law of total probability - Continuous random variable) *Suppose we have a continuous parameter θ in the range* [a,b] *and discrete random data X. Assume θ is itself random with density $f(\theta)$ and that X have likelihood $P(X|\theta)$. In this case, the total probability of X is given by the formula:*

$$P(X) = \int_{\theta} P(X|\theta) f(\theta) d\theta.$$

By Bayes' theorem and the law of total probability, we can retrieve the posterior distribution and therefore compute the conditional expected value of an evolutionary distance.

2.3.2 Probabilites in biological evolution

We assume (a, b) represents an alignment of two protein sequences a and b where a is the the ancestral sequence and b is the descendant sequence. The calculation of the likelihood of (a, b) given an model λ requires us to find the transition probabilities from a to b, as well as the equilibrium frequencies of a.

Let $\mathrm{Q}_\lambda = \{\mathrm{Q}_\lambda\}_{ij}$ denote the instantaneous rate matrix which defines the Markov process in a substitution model λ. Each entry in the matrix corresponds to the rate of change from amino acid i to amino acid j. Let $\overline{F}_\lambda$ denote the vector that consists of all of the equilibrium frequencies of a.

Let $\mathrm{P} = \{\mathrm{P}\}_{ij}$ denote the matrix of transition probabilities for the Markov process. Each entry $p_{i,j}$ in P corresponds to the probability of a site being in state j after time t given that the process started in state i at time 0. To find the matrix P one has:

$$\mathrm{P}(Q_\lambda, t) = e^{tQ_\lambda}.$$

where Q_λ is the instantaneous rate matrix from a substitution model and t is a time unit (Kosiol 2006).

Let a matrix $\mathrm{M} = \{\mathrm{M}\}_{ij}$ denote the number of changes from amino acid i to j in the alignment (a, b). Let $\mathrm{A}_{\lambda,t} = \{\mathrm{A}_{\lambda,t}\}_{ij}$ denote the matrix we will be given if we element-wise raise the elements in the probability matrix P with the elements in matrix M.

The likelihood of (a, b) is retrieved by multiplying the rate of change a_{ij} between sites from matrix $\mathrm{A}_{\lambda,t}$ with the frequency of a particular state f_i from the vector of equilibrium frequencies (Salemi, Vandamme, and Lemey 2009).

The likelihood of the alignment is given by the product of $\overline{F}_\lambda$ and $\mathrm{A}_{\lambda,t}$:

$$P(a, b|t) = \overline{F}_\lambda \cdot \mathrm{A}_{\lambda,t} = \prod_{i,j} f_i a_{ij}.$$

If we assume the prior probability of the time units is $P(t)$, the prior probability of the alignment is $P(a, b)$ and the sequence data has the likelihood $P(a, b|t)$, then by Bayes' rule the posterior probability of t is

$$P(t|a, b) = \frac{P(a, b|t)P(t)}{P(a, b)}.$$

From the law of total probability we can calculate the prior probability of the alignment by

$$P(a, b) = \int_t P(a, b|t)f(t)dt$$

where $f(t)$ is the density function for the time units and $P(a, b|t)$ the likelihood of the alignment.

The distribution of the prior probability of the time units is assumed to be uniform. The reason for this is lack of prior knowledge.

2.3.3 Probability methods

The estimation of the evolutionary distance is usually performed with a maximum likelihood estimation or a maximum a posteriori estimation. A maximum likelihood estimation returns a distance which represents a hypothesis on the evolutionary history, which according to the underlying model, most likely would have given rise to the respective sequence data. By taking the prior probabilities about the distances into account in the hypothesis, we are instead retrieving a maximum a posteriori estimate. Since we assume the distances to be uniformly distributed, the estimates are equal.

For this reason maximum likelihood and maximum a posteriori are both measured by finding the evolutionary distance d, such that d maximizes the likelihood $P(a, b|d)$:

$$\arg\max P(a, b|d).$$

If a maximum likelihood or a maximum a posteriori estimate has been found, we know that the likelihoods of the alignment, with each of the distances in a set used as given data, have been computed. From the set of likelihoods, we can retrieve the posterior probability of an alignment.

We can estimate the posterior distribution of the alignment and compute a maximum likelihood or maximum a posterior estimate during the same iteration. We would then only have to compute one more integral to retrieve the expected value of an evolutionary distance.

2.3.4 Bayesian inference

The evolutionary distance can be estimated using Bayesian inference. In a paper written by Agarwal and States (1996), an estimation of the conditional expected value of evolutionary distances using Bayesian inference is presented.

Let (a, b) represent an alignment of the two protein sequences a and b, let D denote a finite set of evolutionary distances and let d be an evolutionary distance in the set D. The conditional expected value of the distance d given the sequence alignment (a, b) is by Agarwal and States (1996) given by:

$$E(d|a, b) = \sum_{d \in D} d \cdot \Pr(d|a, b).$$

The random variable in the estimate is discrete and is described as all possible time units given our sequence data is (a, b).

The prior probabilities of the distances are assumed to be uniformly distributed and therefore treated as a constant. The prior probability of an evolutionary distance d is therefore given by

$$\Pr(d) = \frac{1}{|D|}.$$

The prior probability that the alignment (a, b) has been generated in the model is, by the law of total probability, given by:

$$\Pr(a,b) = \sum_{d\in D} \Pr(d) \cdot \Pr(a,b|d) = \frac{1}{|\mathrm{D}|} \cdot \sum_{d\in D} \Pr(a,b|d).$$

By following the steps in section 2.3.1, to obtain the posterior $\Pr(d|a, b)$, one has

$$\Pr(d|a,b) = \frac{\Pr(d)\cdot\Pr(a,b|d)}{\Pr(a,b)} = \frac{\frac{1}{|\mathrm{D}|}\cdot\Pr(a,b|d)}{\frac{1}{|\mathrm{D}|}\cdot\sum_{d\in D}\Pr(a,b|d)} = \frac{\Pr(a,b|d)}{\sum_{d\in D}\Pr(a,b|d)}.$$

By that, the conditional expected value by Agarwal and States (1996) is given by:

$$E(d|a,b) = \sum_{d\in D} d \cdot \frac{\Pr(a,b|d)}{\sum_{d\in D}\Pr(a,b|d)}.$$

The expected value in the paper by Agarwal and States (1996) is found from a discrete random variable, meaning the number of evolutionary distances the alignment can have is finite. The evolutionary time units can be defined over an interval of time units and be described as a continuous random variable instead. For that reason a calculation of the conditional expected value of a continuous random variable will be presented.

To be able to use numerical integration to approximate the value of the integral in the continuous case, the infinite set of values in the integral is replaced by a discrete representation. The infinite set of time units are thus transformed at each input of an alignment, to a finite set T. Let t denote an evolutionary time unit in T. The time units are still assumed to be uniformly distributed. We set the probability density function as $f(t) = \dfrac{1}{|T|}$.

From these assumptions, the posterior probability is given by:

$$f(t|a,b) = \frac{P(t)\cdot P(a,b|t)}{P(a,b)} = \frac{P(a,b|t)f(t)}{\int_{t\in\mathrm{T}} P(a,b|t)f(t)dt} = \frac{P(a,b|t)\cdot\frac{1}{|\mathrm{T}|}}{\frac{1}{|\mathrm{T}|}\int_{t\in\mathrm{T}} P(a,b|t)dt} =$$

$$= \frac{P(a,b|t)}{\int_{t\in \mathrm{T}} P(a,b|t)dt}.$$

The conditional expected value is:

$$E(t|a,b) = \int_{t\in \mathrm{T}} t \cdot f(t|a,b)dt = \int_{t\in \mathrm{T}} t \cdot \frac{P(a,b|t)}{\int_{t\in \mathrm{T}} P(a,b|t)dt}dt.$$

3 Problem

The evolutionary distance between two identical sequences will, using maximum likelihood, always be estimated to zero. There exist sequences that are 100 percent identical, but are retrieved from different species, therefore a maximum likelihood estimate does not always have to be accurate (Kumaraswamy and Hatfield 2002). A maximum likelihood estimate does not provide any information regarding the expected values of the evolutionary distances, although, the expected values may reinforce the estimations. In a research by Agarwal and States (1996), the right tools to approximate an expected value of an evolutionary distance using Bayesian inference are presented, but not how an estimation is made efficiently.

3.1 Motivation

The aim of the project is to speed up the integration in the estimation of the conditional expected value of evolutionary distances between sequence of amino acids. We will explore how it is possible to efficiently estimate the expected value by adjusting the discretization of the integral in the computations.

3.1.1 ExpDist

In this report, we propose a new estimator for the evolutionary distances, ExpDist. The data set for the estimator will consist of aligned sequences without any gaps.

3.2 Related work

There exists multiple programs to estimate evolutionary distances using maximum likelihood estimations, two of these are FastMG (Dang et al. 2014) and PhyML (Guindon et al. 2005). These programs do not contain any method to estimate the expected values of the evolutionary distances.

A research made by Agarwal and States (1996) contains tools to estimate the expected value of evolutionary distances. The research does not include any information about how the estimates can be done efficiently.

4 Method

This section contains information about how the biological data for the project was generated, how ExpDist was developed and how the discretization method was improved.

4.1 Generating sequence data

Phylogenetic trees were constructed manually. In each construction, the distances between two species were at least 0.0 and as most 3.0. A phylogenetic tree can be seen in Figure 5.

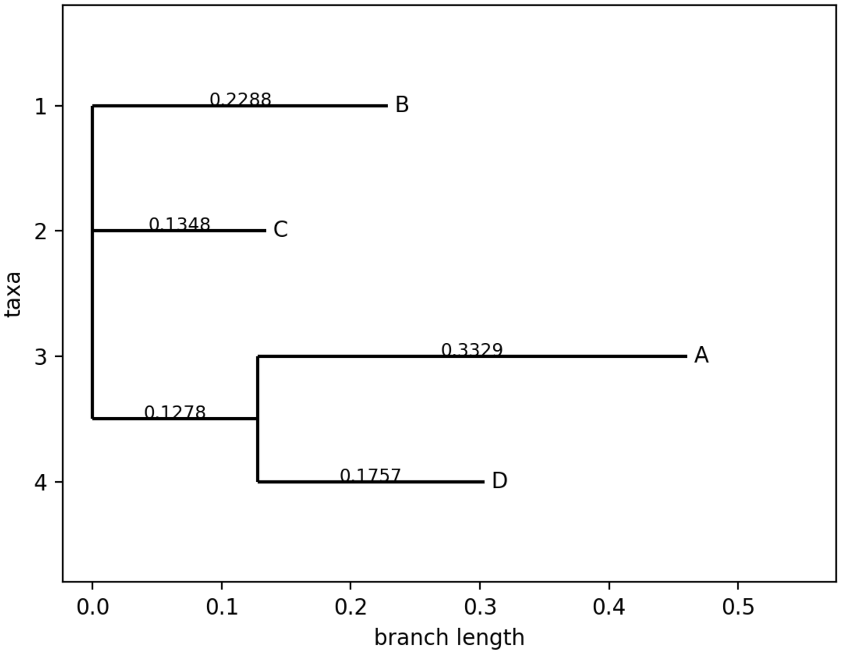
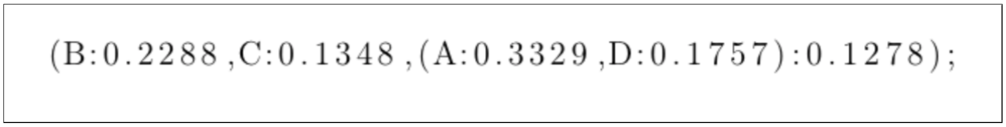


Figure 5: *Phylogenetic tree for the species: A, B, C, D*

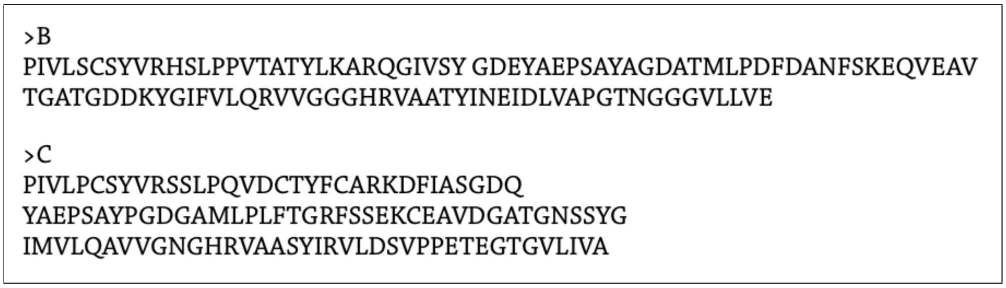
The trees were constructed in a format called Newick format. The Newick format is a phylogenetic tree format for representing trees in a computer-readable form (See Figure 6).



```
(B:0.2288,C:0.1348,(A:0.3329,D:0.1757):0.1278);
```

Figure 6: *The phylogentic tree in Figure 5 represented in Newick format. The species are grouped by parentheses and the length of the branches are included using colons followed by the length of the branch. For example the branch length of A is 0.3329 and the branch length of D is 0.1757. This pair are sibling species and are therefore grouped by parentheses followed by a colon and the branch length which is 0.1228.*

The program Seq-Gen (Rambaut and C. Grass 1997) was used to simulate the evolution of sequences of amino acids along these phylogenetic trees. Seq-Gen takes a phylogenetic tree in the Newick format and a substitution model and returns sequences of amino acids in a format called FASTA format. The FASTA format is a text-based format for representing sequences of amino-acids (See figure 7).



```
>B
PIVLSCSYVRHSLPPVTATYLKARQGIVSY GDEYAEPSAYAGDATMLPDFDANFSKEQVEAV
TGATGDDKYGIFVLQRVVGGGHRVAATYINEIDLVAPGTNGGGVLLVE

>C
PIVLPCSYVRSSLPQVDCTYFCARKDFIASGDQ
YAEPSAYPGDGAMLPLFTGRFSSEKCEAVDGATGNSSYG
IMVLQAVVGNGHRVAASYIRVLDSVPPETEGTGVLIVA
```

Figure 7: *Example of sequences of amino acids in FASTA format generated with Seq-Gen. The data is generated from the phylogenetic tree in Figure 5.*

4.2 Methods for numerical integration

For the integration in the implementation, the Simpson rule and the Trapezoidal rule were implemented. The default method in the implementation of ExpDist is the Simpson rule.

4.3 Probabilites

Rate matrices and equilibrium frequencies from evolutionary models were loaded through a Python module called modelmatcher (Arvestad 2019). The module contains a variety of useful tools for estimations in phylogenetics. In the implementation of ExpDist, substitution models and a method to calculate the transition probability matrix were loaded from the module.

4.4 Discretization

In order to compute the expected value of an evolutionary distance, we used numerical integration. A discretization of the evolutionary distances was done to replace the infinite set of values in the integral by a discrete set of distances, finite in number. In the discretization method, we wanted to construct a set of distances such that it was possible to estimate the expected value fast without loosing important data.

To construct such a set, we used the Poisson correction distance as guidance in the discretization. The Poisson correction distance is fast, but a quite weak estimate of the evolutionary distance. By extending the upper and the lower limit in the intervals with the Poisson correction distance as an initial value, we were able to construct a set of distances which contained the most probable distances.

We created an estimator, ExpDist, where the conditional expected value of the evolutionary distance was implemented as

$$E(t|a,b) = \int_{t=\phi\alpha}^{\phi\beta} t \cdot f(t|a,b)dt = \int_{t=\phi\alpha}^{\phi\beta} t \cdot \frac{P(a,b|t)}{\int_{t=\phi\alpha}^{\phi\beta} P(a,b|t)dt}dt, \qquad (1)$$

where the Poisson correction is denoted by ϕ, the extension of the Poisson correction distance in the lower limit is denoted by α and the extension of Poisson correction distance in the upper limit is denoted by β.

To be able to decide which values of α and β in (1) which gave both fast and accurate outputs, Expdist was compare to a slower but accurate estimator. The estimator was classified as slow but accurate if in the discretization, the interval is between 0 and 3.0 with the step length 0.008. We discovered that the most appropriate values for α and β were changing

Table 1: *The evolutionary distances between the species in the trees.*

Tree	Evolutionary distance between species
W	0.8 - 1.5
X	0.0 - 0.3
Y	0.3 - 1.0
W	2.4 - 3.0

depending on the Poisson correction distance. Therefore the discretization was split into different cases which did depend on the Poisson correction distance. Testing was then made for each of these cases, in order to decide the most appropriate values for α and β.

4.4.1 Test cases

Four different phylogenetic trees were constructed, each tree with three species. The trees have species with different evolutionary distances. The trees and their different interval of evolutionary distances can be seen in Table 1. The main idea was from these distances, be able to adjust the Poisson correction distance in the discretization.

Sequences were generated from the trees using PAM, WAG, LG and JTT as substitutions models. The sequences were generated 35 times per tree and substitution model. Therefore a total of 420 alignments per model were constructed.

The same model that was used as input for Seq-Gen to generate the sequences was used to estimate the expected value of the evolutionary distance between the sequences. The regular length of the sequences was 110.

The estimator ExpDist and all test cases is available at a repository at GitHub (Radenholt 2020).

4.4.2 Discretization method

We recognized that, the expected value estimated with the slow and accurate estimator, did in in some cases differ from the correct answer. It would be improper to request ExpDist to approach the correct answer, if the correct answer is not approached with the slow, accurate estimator. For this reason, the expected value estimated with ExpDist was both compared to the correct answer and to the expected value estimated with the slow estimator.

The idea was that the expected value, computed from the set of distances constructed in the discretization, had to fulfill the following conditions. Either we wanted the expected value computed with ExpDist to be as most 10 percent from the expected value estimated by the slow estimator or the correct answer to be in at most a standard deviation from the expected value computed with ExpDist. The standard deviation σ was calculated as

$$\sigma = \sqrt{E((t^2|a,b)) - E(t|a,b)^2} = \sqrt{\int_{t=\phi\alpha}^{\phi\beta} (t^2 - t) \frac{P(a,b|t)}{\int_{t=\phi\alpha}^{\phi\beta} P(a,b|t)dt} dt}.$$

During the testing, different values for α and β, the step-size Δt were tried.

The discretization was divided into four different cases. In these cases we let the values of α, β and Δt in the interval be dependent of the Poisson correction distance of the alignments. The cases can be seen in Table 2.

Table 2: *The different cases for the construction of the set of discretization points.*

Case	**Poisson correction distance ϕ**	**Lower limit**	**Upper limit**	**Δt**
1	ϕ <0.1	0.0	0.2	0.008
2	ϕ <0.4	0.4ϕ	1.45ϕ	0.01
3	$0.9 < \phi < 1.35$	0.9ϕ	1.9ϕ	0.15
4	ϕ >1.35	0.8ϕ	3.0	0.4

5 Result and analysis

ExpDist demonstrates that it is possible to efficiently estimate the expected value of evolutionary distances. To speed up the integration in the computations, without losing important data, the Poisson correction distances can be used as a quick and easy guidance in the construction of the set of discretization points.

In the following figures, ExpDist is compared to the slow and accurate estimator. In Figure 8 we can see how close the estimates are to the correct answers, the difference in the discretization points between the estimators and the standard deviations of the expected values approximated with ExpDist. In Figure 9 and Figure 10 we can see the posterior distribution in ExpDist compared to the posterior distribution in the slow estimator.

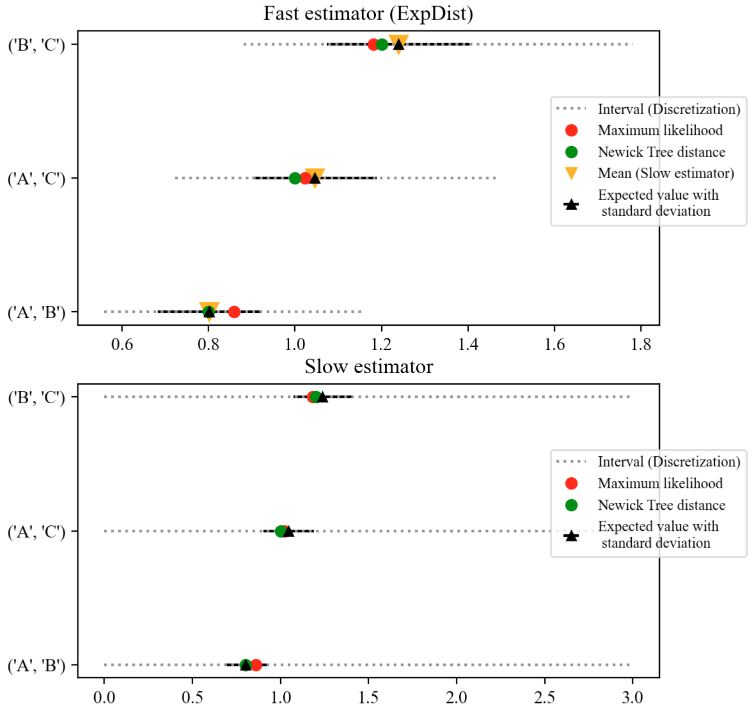


Figure 8: *The plots are from estimates of evolutionary distances between sequences from test case 'W' (For details, see section 4.4.1). The estimates with ExpDist (first plot) can be compared to the estimates with the slow estimator (second plot). The x-axis represents the evolutionary distances. The y-axis represents the pair of sequences which been used as input. From these plots it is possible to read and compare the expected value, the maximum likelihood estimate, the correct answer and the specific discretization of the evolutionary distances.*

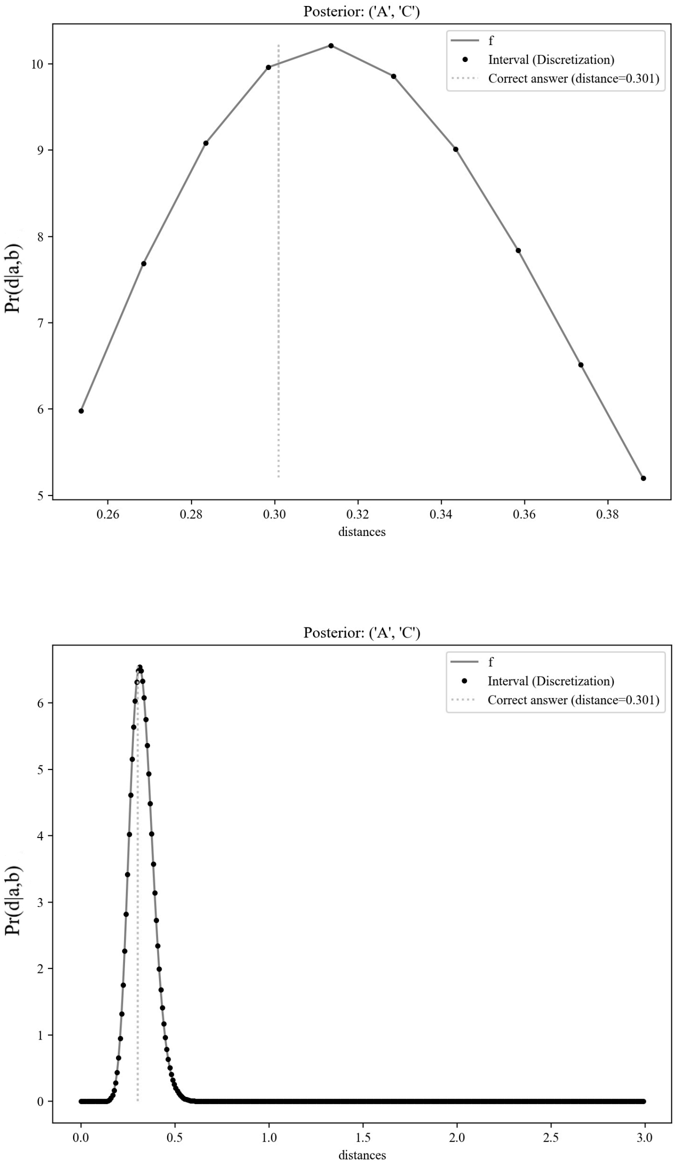


Figure 9: *The posterior distribution in ExpDist compared to the slow estimator. The first plot demonstrates the posterior distribution in ExpDist. The second plot demonstrates the posterior distribution in the slow estimator for the same test case. The x-axis represents the evolutionary distances and the y-axis represents the posterior probability. The dots visualize the specific discretization of evolutionary distances and the vertical line represents the correct answer. In the first plot, we can see that the discretization points are concentrated near the Poisson correction distance and hit maximum likelihood. The posterior distribution is retrieved from the estimate of the expected value of the evolutionary distance between sequences from test case 'X' (For details, see section 4.4.1).*

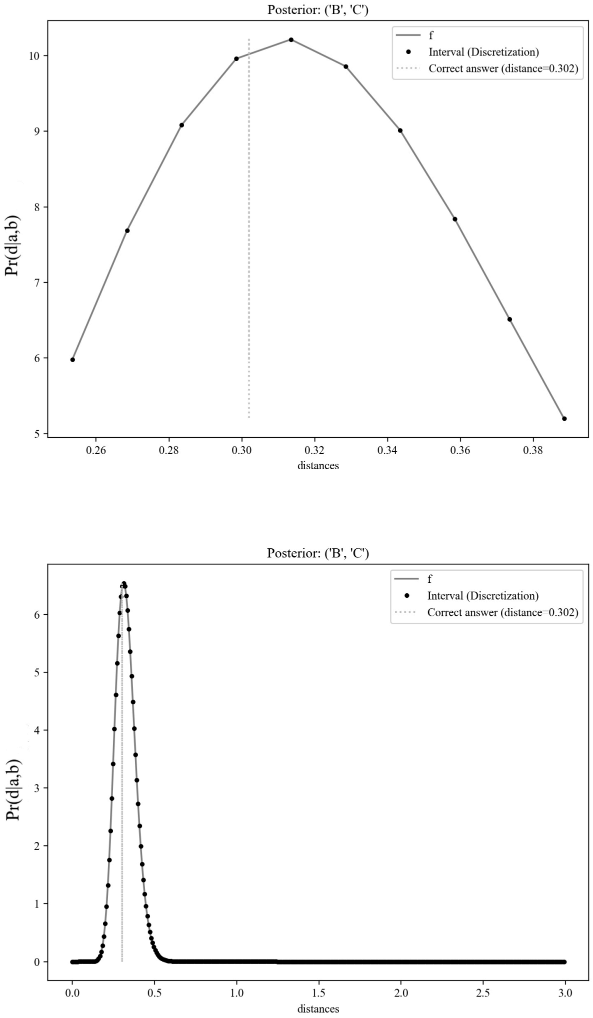


Figure 10: *The posterior distribution in ExpDist compared to the slow estimator. The first plot demonstrates the posterior distribution in ExpDist. The second plot demonstrates the posterior distribution in the slow estimator for the same test case. The x-axis represents the evolutionary distances and the y-axis represents the posterior probability. The dots visualize the specific discretization of evolutionary distances and the vertical line represents the correct answer. In the first plot, we can see that the discretization points are concentrated near the Poisson correction distance and hit maximum likelihood. The posterior distribution is retrieved from the estimate of the expected value of the evolutionary distance between sequences from test case 'X' (For details, see section 4.4.1).*

The size of the set of discretization points for the slow estimator is 375. The actual minimum and maximum number of discretization points in ExpDist are found in Table 3.

Table 3: *The minimum and the maximum size of the sets in the discretization in ExpDist. The size of the set is dependent on the Poisson correction distance ϕ.*

Poisson correction distance ϕ	**Lower limit**	**Upper limit**	**min |T|**	**max |T|**
ϕ <0.1	0.0	0.2	25	25
ϕ <0.4	0.4ϕ	1.45ϕ	3	13
0.9 < ϕ < 1.35	0.9ϕ	1.9ϕ	6	9
ϕ >1.35	0.95ϕ	3.0	19	-

5.1 Limitations in the implementation

In the estimations for the expected value of the evolutionary distance between longer sequences, there exists cancellations effects. Consequently, the length of the sequences used as input for ExpDist is limited. Details about how it is possible to approach this problem are presented at section 6.1 in this report.

6 Conclusions

In the study of protein there exist weaknesses in the most commonly methods which are used to predict the evolution. The expected value of an evolutionary distance has been shown to robust the estimation and in this report we have developed an efficient estimator for the expected values of evolutionary distances.

The result shows that if we, in the discretization, let the upper limit, the lower limit and the change Δt in the intervals be dependent of the Poisson correction distance, the number of discretization points can be reduced without losing important data.

During the implementation, a problem regarding the length of the sequences used as input for ExpDist occurred. Fortunately, possible approaches for the problem are presented for further research.

6.1 Suggestion for further research

Because of cancellations effects in the estimator, the expected values of the evolutionary distances for sequences with just over 110 sites could not be approximated. For this reason, suggestions to implement the methods using logarithms have been made (Baldi and Brunak 2001).

In this report and in the implementation of ExpDist, the discretization has been based on four cases. If one continues to investigate how the selection of the set of points can be adjusted from particular proportion of Poisson correction distances, there are opportunities to improve the estimation of the expected value of evolutionary distances to a greater extent.

References

Agarwal, Pankaj and David.J States (1996). "A Bayesian evolutionary distance for Parametrically Aligned Sequences". In: URL: http://www.stateslab.org/publications/agarwal%20bayesian%20dist%20jcb%201996.pdf.

Arvestad, Lars (2019). *modelmatcher: Rapid identification of evolutionary models*. URL: https://pypi.org/project/modelmatcher/.

Baldi, Pierre and Søren Brunak (2001). *Bioinformatics: The machine learning approach*. MIT press.

Baum, David (2008). "Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups". In: *Nature Education* 1 (1), p. 190. URL: https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/.

Dang, Cuong Cao et al. (2014). "FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets". In: *BMC Bioinformatics* 15 (1), p. 341. DOI: https://doi.org/10.1186/1471-2105-15-341.

Dayhoff, M.O., R. Schwartz, and B.C Orcutt (1967). "A model of Evolutionary Change in Proteins: Atlas of protein sequence and structure". In: *Systematic Biology* 16 (3), pp. 262–263. DOI: https://doi.org/10.2307/2412074.

Garrett, Reginald H. and Charles M. Grisham (2014). *Biochemistry*. Cengage Learning.

Guindon, Stephanie et al. (2005). "PHYML: a web server for fast maximum likelihood-based phylogenetic inference". In: *Nucleic Acids Research* 33 (2), pp. 557–559. DOI: https://doi.org/10.1093/nar/gki352.

Jones, David T, William R Taylor, and Janet M Thornton (1992). "The rapid generation of mutation data matrices from protein sequences". In:

Bioinformatics 8 (3), pp. 275–282. DOI: https://doi.org/10.1093/bioinformatics/8.3.275.

Kosiol, Carolin (2006). “Markov Models for Protein Sequence Evolution”. In: URL: https://www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/documents/phdtheses/carolinkosiolthesis.pdf.

Kumaraswamy, Easwari and Dolph L. Hatfield (2002). *Methods in Enzymology*. Academic Press.

Le, SQ and O. Gascuel (2008). “An improved general amino acid replacement matrix.” In: *Molecular Biology and Evolution* 25 (7), pp. 1307–1320. DOI: https://doi.org/10.1093/molbev/msn067.

Nei, Masatoshi and Jianzhi Zhang (2005). “Evolutionary Distance: Estimation”. In: DOI: https://doi.org/10.1038/npg.els.0005108.

Radenholt, Julia (2020). *Repository: ExpDist*. URL: https://github.com/juliaradenholt/expDist/.

Rambaut, Andrew and Nicholas C. Grass (1997). “Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees”. In: *Bioinformatics* 13 (3), pp. 235–238. DOI: https://doi.org/10.1093/bioinformatics/13.3.235.

Salemi, Marco, Anne-Mieke Vandamme, and Philippe Lemey (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.

Whelan, Simon and Nick Goldman (2001). “A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach”. In: *Molecular Biology and Evolution* 18 (5), pp. 691–699. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a003851.