

EvoVGM: a Deep Variational Generative Model for Evolutionary Parameter Estimation

Amine M. Remita

remita.amine@courrier.uqam.ca
Department of Computer Science
Université du Québec à Montréal
Montreal, Quebec, Canada

Abdoulaye Baniré Diallo

diallo.abdoulaye@uqam.ca
Department of Computer Science
Université du Québec à Montréal
Montreal, Quebec, Canada

ABSTRACT

Most evolutionary-oriented deep generative models do not explicitly consider the underlying evolutionary dynamics of biological sequences as it is performed within the Bayesian phylogenetic inference framework. In this study, we propose a method for a deep variational Bayesian generative model (EvoVGM) that jointly approximates the true posterior of local evolutionary parameters and generates sequence alignments. Moreover, it is instantiated and tuned for continuous-time Markov chain substitution models such as JC69, K80 and GTR. We train the model via a low-variance stochastic estimator and a gradient ascent algorithm. Here, we analyze the consistency and effectiveness of EvoVGM on synthetic sequence alignments simulated with several evolutionary scenarios and different sizes. Finally, we highlight the robustness of a fine-tuned EvoVGM model using a sequence alignment of gene S of coronaviruses.

CCS CONCEPTS

• **Applied computing** → **Molecular evolution**; *Molecular sequence analysis*; • **Mathematics of computing** → *Bayesian computation*; **Variational methods**; • **Computing methodologies** → *Learning latent representations*; **Latent variable models**; *Neural networks*.

KEYWORDS

Variational Generative Model, Evolutionary model, Substitution model, Variational inference, Latent variables, Deep neural networks, EvoVGM

ACM Reference Format:

Amine M. Remita and Abdoulaye Baniré Diallo. 2022. EvoVGM: a Deep Variational Generative Model for Evolutionary Parameter Estimation. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22)*, August 7–10, 2022, Northbrook, IL, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3535508.3545563>

1 INTRODUCTION

In systematics and evolutionary biology, probabilistic evolutionary models are extensively used to study unseen and complex historical events affecting the genomes of a set of taxa during a period of

time (i.e., recombination, horizontal gene transfer and selective pressure). Their ability to detect evolutionary events and measure their parameters using biological sequences has enabled valuable applications in population genetics [13], medicine [33] and epidemiology [4, 5]. These models allow the estimation of probabilities of certain types of mutations such as substitutions [11, 27], indels [3] and genome rearrangements [24]. Main approaches supporting evolutionary studies, such as phylogenetics, implement evolutionary models with Markovian properties [27].

Typically, evolutionary parameters of these models are jointly represented with different types of high-dimensional variables (discrete and continuous), inducing a computationally intractable joint posterior. Bayesian phylogenetic approaches provide methods to efficiently approximate the intractable joint posterior and quantify the uncertainty in the estimation of the parameters [9, 32]. They mainly implement random-walk Markov Chain Monte Carlo (MCMC) algorithms, which can converge to an accurate posterior but with a considerable cost. Furthermore, they are prone to limitations due to the complexity of the posterior [29], their dependence on initialization and proposal distribution parameters, and their sensitivity to the prior distributions [8]. Recently, variational inference (VI) has sparked interest in phylogenetics as a robust alternative to approximate the intractable posterior by relying on fast optimization methods [2, 7, 34, 35]. VI finds an optimal candidate from a space of tractable distributions that minimizes the Kullback-Leibler (KL) divergence to the exact posterior [1, 10]. It inherently bounds the intractable marginal likelihood of the observed data. Moreover, VI is also used in building deep generative models [16, 21]. However, contrary to Bayesian phylogenetic inference frameworks, most evolutionary-oriented deep generative models do not explicitly consider the underlying evolutionary dynamics of the biological sequences [19, 22, 28].

Here, we propose EvoVGM, a deep variational generative model that simultaneously estimates local evolutionary parameters and generates nucleotide sequence data. Like phylogenetic inference, we explicitly integrate a continuous-time Markov chain substitution model into the generative model. The model is trained in an unsupervised manner following the evolutionary model constraints.

2 BACKGROUND

2.1 Notation

The observed data \mathbf{X} is an alignment of M character sequences with length N , where $\mathbf{X} \in \mathcal{A}^{M \times N}$. In our case, the alphabet of characters $\mathcal{A} = \{A, G, C, T\}$ is a set of nucleotides. x_n^m is the character in the m^{th} sequence (x^m) and at the n^{th} site (x_n) of the alignment. Here, we assume that each alignment \mathbf{X} has a hidden ancestral state

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
BCB '22, August 7–10, 2022, Northbrook, IL, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9386-7/22/08.
<https://doi.org/10.1145/3535508.3545563>

sequence $a \in \mathcal{A}^N$. We take the hypothesis that each ancestral state a_n has evolved independently from the other states $\{a_i; i \neq n\}$ to an extant character x_n^m over an evolutionary time expressed as a branch length t and following a substitution model defined by a set of parameters ψ . In a Bayesian framework, we seek representations allowing to model uncertainty on the quantity and the composition of different entities. We consider the observable characters (x_n^m) and the ancestral states (a_n) as random variables (noted in bold, unlike scalar values) and represent them by categorical distributions over \mathcal{A} . Also, branch lengths (t^m) and substitution model parameters (ψ) will be modelled as random variables and will be represented by suitable distributions.

2.2 Markov Chain Models of Character Substitution

The evolution of a character is measured by the number of hidden substitutions that undergoes over time. To estimate this quantity, we assume that the process of evolution follows a continuous-time Markov chain model whose states belong to \mathcal{A} . The model is parameterized by a rate matrix Q and relative frequencies π of characters at equilibrium. Each element of the matrix q_{ij} ($i \neq j$) defines the instantaneous substitution rate of character i changing into character j . The diagonal elements q_{ii} are set up in a way that each row sums to 0. Q is scaled by a factor μ , so that the time t will be measured in the expected number of substitutions per site and the average rate of substitution at equilibrium will be 1. We use time-reversible Markov chain models assuming the amount of changes from one character to another is the same in both ways. For nucleotide substitution time-reversible models, the equation of Q is

$$Q = \begin{pmatrix} \cdot & a\pi_G & b\pi_C & c\pi_T \\ a\pi_A & \cdot & d\pi_C & e\pi_T \\ b\pi_A & d\pi_G & \cdot & f\pi_T \\ c\pi_A & e\pi_G & f\pi_C & \cdot \end{pmatrix} \mu,$$

where a, b, c, d, e, f are the set of relative substitution rate parameters ρ , and $\pi_A + \pi_G + \pi_C + \pi_T = 1$ are the relative frequencies π . Once Q is estimated we can compute the probability transition matrix P over an evolutionary time t as $P(t) = \exp(Q t)$. The matrix exponential is computed using spectral decomposition of Q as it is reversible (see [17] and [31] for more details).

Several substitution models could be generated depending on the constraints placed on the set of parameters $\psi = \{\rho, \pi\}$. The simplest model is JC69 with equal substitution rates and uniform relative frequencies [11]. The K80 model defines uniform frequencies like JC69, but it differentiates between the two types of substitution rates corresponding to transitions ($\alpha = a = f$) and transversions ($\beta = b = c = d = e$) [14]. Usually, K80 is parameterized by the transition/transversion rate ratio $\kappa = \alpha/\beta$. Finally, the general time-reversible (GTR) model sets all the parameters ψ free [27, 30].

2.3 Evolutionary Posterior

Along with a and t variables, we consider the parameters of the Markov chain model ψ as latent (hidden) variables to be inferred from the observed data X . Assuming an independent evolution of the sites in an alignment [6], the marginal likelihood of the data X factorizes into $p(X) = \prod_{n=1}^N p(x_n)$. The inference of the latent

variables for each site x_n requires the computation of the evolutionary joint posterior $p(a_n, t, \psi | x_n)$. The evolutionary posterior is calculated according to Bayes' theorem:

$$p(a_n, t, \psi | x_n) = \frac{p(x_n, a_n, t, \psi)}{p(x_n)}, \quad (1)$$

which exposes the joint density of the observable variable and the latent variables $p(x_n, a_n, t, \psi)$, and the marginal likelihood $p(x_n)$. The former is factorized as a product of the joint prior density of the latent variables $p(a_n, t, \psi)$ and the likelihood $p(x_n | a_n, t, \psi)$. The latter is calculated by marginalizing over the values of all the latent variables as $\iiint p(a_n, t, \psi) p(x_n | a_n, t, \psi) da_n dt d\psi$. The computation of the evolutionary joint posterior density is computationally intractable as it depends on the evaluation of $p(x_n)$, which is intractable due to the integrals in its marginalization. We show in the next section strategies to determine each term in the equation 1.

3 PROPOSED EVOLUTIONARY MODEL

In this section, we describe a deep variational generative model that simultaneously estimates local evolutionary biological parameters and generates nucleotide sequence data. Similar to deep variational-based generative models [16, 21], the proposed model architecture consists of two main sub-models: 1) a set of deep variational encoders that infers the parameters of evolutionary-latent-variable distributions and allows sampling, and 2) a generating model that computes probability transition matrices from sampled latent variables and generates a distribution of sequence alignments from reconstructed ancestral states (see Figure 1).

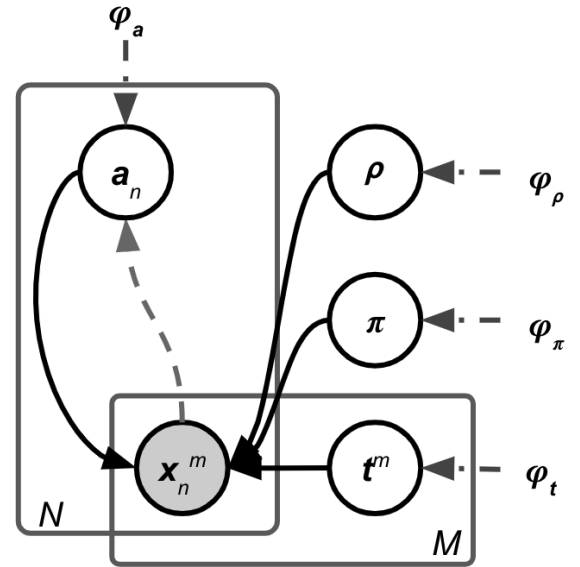


Figure 1: Graphical illustration of the inference (dashed gray lines) and the generation (solid lines) processes of the GTR-based variational generative model. Gray circles represent the observed variables. Blank circles represent the latent variables. $\{\varphi_a, \varphi_t, \varphi_\rho, \varphi_\pi\}$ is the set of hyper-parameters of the prior densities.

3.1 Variational Inference of the Joint Posterior

We use mean-field variational inference to approximate the true joint posterior probability distribution by a new probability distribution $q_\phi(\mathbf{a}_n, \mathbf{t}, \psi | \mathbf{x}_n)$ [10, 16, 21]. We model each latent variable by an independent approximate distribution whose parameters will be inferred using a non-linear transformation either of \mathbf{x}_n , or an independent, fixed random noise ζ . The non-linear transformations are implemented using deep neural networks (NeuralNet) parameterized by a set of independent and adaptable variational parameters $\phi = \{\phi_a, \phi_t, \phi_\psi\}$.

For each sequence x^m , we infer and sample an evolutionary time variable \mathbf{t}^m . We model its approximate density $q_{\phi_t}(\mathbf{t}^m)$ using a gamma distribution to ensure the positiveness of the samples. The parameters of the distribution (shape and rate) are produced by a non-linear transformation on uniform noise ζ_t as follows:

$$q_{\phi_t}(\mathbf{t}^m) = \text{Gamma}(\mathbf{t}^m; \text{NeuralNet}(\zeta_t; \phi_t)).$$

Next, we infer and sample the latent variables of the Markov chain model parameters ψ with independent approximate densities $q_{\phi_\psi}(\psi)$. The JC69 model does not have any free parameters to be estimated, so $\psi = \emptyset$. For the K80 model, we infer the latent variable of the transition/transversion rate ratio (κ) using a gamma-based approximate distribution ($q_{\phi_\kappa}(\kappa)$) to ensure the positiveness of the samples. Its local parameters are produced by a neural network on uniform noise ζ_κ as follows:

$$q_{\phi_\kappa}(\kappa) = \text{Gamma}(\kappa; \text{NeuralNet}(\zeta_\kappa; \phi_\kappa)).$$

In the case of the GTR model, we model the variational densities of the substitution rate parameters (ρ) and the relative frequencies (π) using Dirichlet distributions. This ensures that the sum of the sampled values is equal to one. Their concentrations are generated by a set of independent neural networks on uniform noises ζ_ρ and ζ_π , respectively:

$$q_{\phi_\rho}(\rho) = \text{Dirichlet}(\rho; \text{NeuralNet}(\zeta_\rho; \phi_\rho)),$$

$$q_{\phi_\pi}(\pi) = \text{Dirichlet}(\pi; \text{NeuralNet}(\zeta_\pi; \phi_\pi)).$$

Lastly, for each site x_n , an ancestral variable \mathbf{a}_n is inferred and sampled with an approximate density $q_{\phi_a}(\mathbf{a}_n | \mathbf{x}_n)$ represented by a categorical distribution over the $(|\mathcal{A}| - 1)$ -simplex as follows:

$$q_{\phi_a}(\mathbf{a}_n | \mathbf{x}_n) = \text{Categorical}(\mathbf{a}_n; \text{NeuralNet}(\mathbf{x}_n; \phi_a)).$$

We apply a non-linear transformation on \mathbf{x}_n to produce the local parameters of $q_{\phi_a}(\mathbf{a}_n | \mathbf{x}_n)$, which are a set of $|\mathcal{A}|$ probabilities that sum to one. Using a mean-field variational inference approach, the approximate joint posterior factorizes into:

$$q_\phi(\mathbf{a}_n, \mathbf{t}, \psi | \mathbf{x}_n) = q_{\phi_a}(\mathbf{a}_n | \mathbf{x}_n) \prod_{m=1}^M q_{\phi_t}(\mathbf{t}^m) q_{\phi_\psi}(\psi). \quad (2)$$

3.2 Generating Model Computation

The generating model is represented by the joint density $p(\mathbf{x}_n, \mathbf{a}_n, \mathbf{t}, \psi) = p(\mathbf{a}_n, \mathbf{t}, \psi) p(\mathbf{x}_n | \mathbf{a}_n, \mathbf{t}, \psi)$, which is parameterized only by the local latent variables. We use independent prior densities for the latent variables, so $p(\mathbf{a}_n, \mathbf{t}, \psi) = p(\mathbf{a}) p(\mathbf{t}) p(\psi)$. To ease the computation, we apply for each prior density the same distribution type as its corresponding approximate posterior density and determine its hyper-parameters ϕ . Moreover, for each nucleotide x_n^m , we use the

probability transition matrix $\mathbf{P}(\mathbf{t}^m)$ to define the likelihood function, which is the probability of evolving a character \mathbf{a}_n into \mathbf{x}_n^m during a time \mathbf{t}^m , as:

$$\begin{aligned} \hat{\mathbf{x}}_n^m &= \mathbf{a}_n \times \mathbf{P}(\mathbf{t}^m; \psi), \\ p(\mathbf{x}_n^m | \mathbf{a}_n, \mathbf{t}^m, \psi) &= \text{Categorical}(\mathbf{x}_n^m; \hat{\mathbf{x}}_n^m). \end{aligned} \quad (3)$$

The likelihood of a site x_n is computed following a pre-order traversal. We call it a top-down likelihood since it includes the sampled ancestral states in its estimation. It is different from the likelihood computed in a phylogeny, which is based on a post-order traversal [6] and does not include sampled ancestral states. Finally, the joint density is

$$p(\mathbf{x}_n, \mathbf{a}_n, \mathbf{t}, \psi) = p(\mathbf{a}) p(\mathbf{t}) p(\psi) \prod_{m=1}^M p(\mathbf{x}_n^m | \mathbf{a}_n, \mathbf{t}^m, \psi). \quad (4)$$

3.3 Stochastic Estimator and Learning Algorithm

Variational inference allows us to form a lower bound on the marginal likelihood of each site x_n as $\log p(\mathbf{x}_n) \geq \mathcal{L}_n(\phi, \mathbf{x}_n)$, where \mathcal{L}_n is the evidence lower bound (ELBO) [1, 10]. Putting together equations 1, 2 and 4, we can derive the equation of the multi-sample estimator of the EvoVGM model as follows:

$$\begin{aligned} \mathcal{L}_n(\phi, \mathbf{x}_n) &= \left(\frac{1}{L} \sum_{l=1}^L \sum_{m=1}^M \log p(\mathbf{x}_n^m | \mathbf{a}_n^l, \mathbf{t}^{m,l}, \psi^l) \right) \\ &\quad - \alpha_{\text{KL}} \left(\text{KL}(q_{\phi_a}(\mathbf{a}_n | \mathbf{x}_n) \parallel p(\mathbf{a})) + \sum_{m=1}^M \text{KL}(q_{\phi_t}(\mathbf{t}^m) \parallel p(\mathbf{t})) \right. \\ &\quad \left. + \text{KL}(q_{\phi_\psi}(\psi) \parallel p(\psi)) \right), \end{aligned} \quad (5)$$

where L is the sampling size, $\text{KL}(\cdot \parallel \cdot)$ is the Kullback–Leibler divergence, and α_{KL} is a regularization coefficient (see the development of this equation in A.1). This estimator is computationally tractable because it is independent of the direct evaluation of the true joint posterior. To maximize the ELBO and learn the global variational

Algorithm 1 Learning algorithm for EvoVGM

Input: Alignment \mathbf{X} of M sequences with length N
 $\phi_a, \phi_t, \phi_\psi \leftarrow$ initialize global variational parameters
for $i \in [1 \dots \text{max_iter}]$ **do**
 $\mathbf{t}^m \leftarrow$ Sample $M \times L$ branch latent variables (ϕ_t)
 $\psi \leftarrow$ Sample L evolutionary latent variables (ϕ_ψ)
 $\mathbf{P}^m \leftarrow$ Compute $M \times L$ probability transition matrices (\mathbf{t}^m, ψ)
for $n \in [1 \dots N]$ **do**
 $\mathbf{a}_n \leftarrow$ Sample L ancestor latent variable ($\mathbf{x}_n; \phi_a$)
 $\hat{\mathbf{x}}_n \leftarrow$ Generate $M \times L$ nucleotides ($\mathbf{a}_n, \mathbf{P}^m$)
 $\mathcal{L}_n \leftarrow$ Compute ELBO according to the equation 5
 $\mathcal{L} += \mathcal{L}_n$
end for
 $\mathbf{g} \leftarrow$ Compute gradients of total ELBO (\mathcal{L})
 $\phi_a, \phi_t, \phi_\psi \leftarrow$ Update parameters (\mathbf{g}) with gradient ascent optimizer
end for

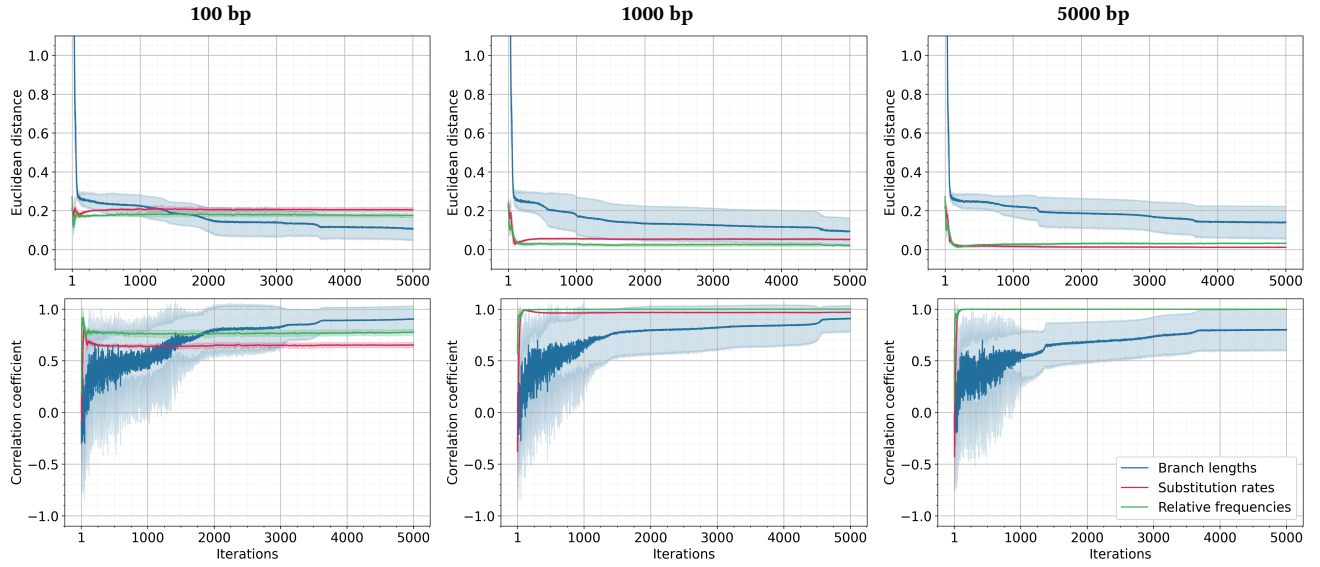


Figure 2: Performance of EvoVGM_GTR model over several lengths of validation sequence alignments (100 bp, 1000 bp and 5000 bp). The GTR substitution model was used to simulate training and validation alignments of five sequences. The results are computed and averaged from fitting and running EvoVGM_GTR ten times. They are reported in terms of Euclidean distance and Pearson correlation coefficient of estimated and actual evolutionary parameters during fitting.

parameters ϕ , EvoVGM estimates and backpropagates the gradients for the whole data X using the reparameterization trick [16] and a gradient ascent optimizer. The algorithm of EvoVGM is detailed in Algorithm 1. It is implemented in Pytorch [20] and its open-source code is available at <https://github.com/maremita/evoVGM>.

4 EXPERIMENTS

The evaluation of the proposed Bayesian variational model to estimate evolutionary parameters and generate sequence alignments is oriented towards assessing its consistency, effectiveness, and understanding its behavior during the training using simulated sequence alignments. Moreover, we highlight the robustness of a fine-tuned EvoVGM model using a sequence alignment of gene S of coronaviruses.

Sequence Alignment Simulation. We used Pyvolve [25] to simulate the evolution of different sequence alignments with a site-wise homogeneity model and a combination of substitution models (JC69, K80 and GTR), the number of sequences (3, 4 and 5) and alignment lengths (100 bp, 1000 bp and 5000 bp). A site-wise homogeneity model evolves sequences from a root sequence with the same substitution model over lineages and with the same branch lengths for nucleotides. The sequence alignments used in the training step of the EvoVGM models were simulated with different random seeds from those used in the validation step but with the same array of evolutionary parameters.

Performance metrics. We report the results of the performance of EvoVGM models in terms of the ELBO, the log likelihood (LogL), and the KL divergence between the approximate densities and the priors (KL_qp) on the training and the validation sequence alignments. To

assess the accuracy of the estimation of the evolutionary parameters, we compute the Euclidean distance and the Pearson correlation coefficient between their arrays and those of actual parameters used in the simulation of the alignments.

4.1 Hyper-Parameters Fine-tuning

First, we assessed the effects of different hyper-parameters on the convergence and the accuracy of the EvoVGM models to approximate the true distributions of the evolutionary parameters. For each hyper-parameters setting, the model was trained ten times, using different weight initialization, on the same alignment of five 5000-bp sequences. Based on the results of a grid search with different combinations of hyper-parameters, we defined the default components of the EvoVGM models including a set of one-hidden-layer variational encoders with a hidden size of 32. We set uniform hyper-parameters on the prior densities of the ancestral states, the substitution rates and the relative frequencies, and we placed independent gamma priors on the branch lengths with a prior expectation of 0.1. Moreover, we used a 100-sample EvoVGM estimator with a regularization coefficient α_{KL} equals 10^{-3} and Adam optimizer [15] for stochastic gradient ascent with a learning rate of 0.005.

Figures A.1, A.2, A.3 and A.4, in the Supplemental Results section, highlight the convergence and the performance of EvoVGM_GTR model (implementing the GTR substitution model) trained with multiple values of the coefficient α_{KL} , the size of the hidden layers, the sampling size, and the learning rate, respectively. In general, EvoVGM_GTR models converge faster when the α_{KL} coefficient is lower, and the number of hidden layers and the learning rate are larger. The sample size does not affect the overall convergence.

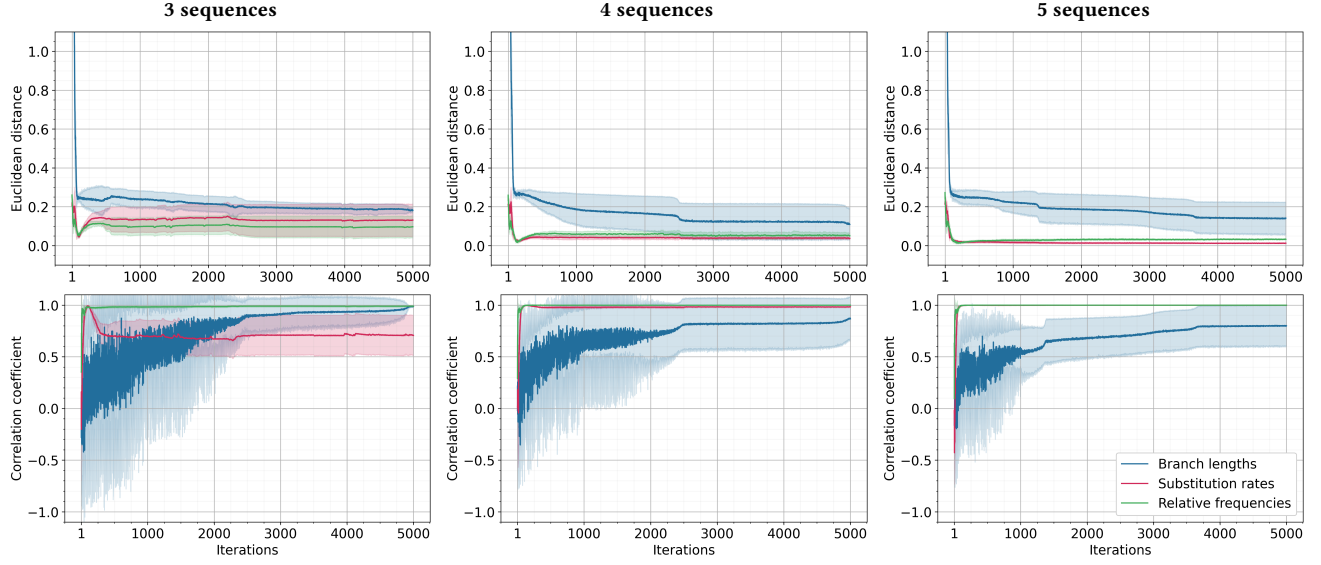


Figure 3: Performance of EvoVGM_GTR model over several numbers of validation sequences (3, 4 and 5). The GTR substitution model was used to simulate training and validation alignments with a length of 5000 bp. The results are computed and averaged from fitting and running EvoVGM_GTR ten times. They are reported in terms of Euclidean distance and Pearson correlation coefficient of estimated and actual evolutionary parameters during fitting.

Table 1: Performance of trained EvoVGM_GTR model using validation alignments simulated with different sizes. The GTR substitution model was used to simulate training and validation alignments. The results are reported in terms of Euclidean distance (DIST) and Pearson correlation coefficient (CORR and PVAL) of estimated and actual evolutionary parameters.

$N \rightarrow$		100			1000			5000		
	M	DIST	CORR	PVAL	DIST	CORR	PVAL	DIST	CORR	PVAL
BRANCH LENGTHS	3	0.300	0.984	0.114	0.090	0.980	0.126	0.097	0.986	0.107
	4	0.076	0.994	0.006	0.086	0.998	0.002	0.085	0.992	0.008
	5	0.081	0.986	0.002	0.069	0.995	0.000	0.116	0.962	0.009
SUBSTITUTION RATES	3	0.621	0.103	0.846	0.177	0.668	0.147	0.129	0.784	0.065
	4	0.305	0.472	0.344	0.114	0.864	0.027	0.036	0.985	0.000
	5	0.206	0.652	0.160	0.053	0.968	0.002	0.012	0.998	0.000
RELATIVE FREQUENCIES	3	0.190	0.941	0.059	0.084	0.991	0.009	0.095	0.992	0.008
	4	0.125	0.891	0.109	0.090	0.996	0.004	0.050	0.999	0.001
	5	0.176	0.775	0.225	0.022	1.000	0.000	0.033	0.999	0.001

However, a small sample size induces a substantial variance in the estimator.

4.2 Assessing Consistency on Simulated Data

Next, we analyzed the consistency and the effectiveness of the EvoVGM_GTR model on sequence alignments simulated with different sizes in terms of length and number of sequences. We built the model using the same default configuration and the same hyperparameters defined in the previous analysis. Figure 2 shows the accuracy of EvoVGM_GTR during its training in terms of the Euclidean distance and the correlation coefficient of the estimated evolutionary parameters using validation sequence alignments of

five sequences with lengths of 100 bp, 1000 bp and 5000 bp. Conversely, Figure 3 shows the accuracy of the model in approximating the parameters on validation sequence alignments of three, four and five sequences with a length of 5000 bp. Usually, the parameter approximation is improved when the number of sequences is higher, and the alignments are longer. On the one hand, branch lengths suffer from high variance estimation and slow evolution at the beginning of the training across all datasets. However, its variance and accuracy improve with more training iterations. On the other hand, the estimation of the substitution rates and the relative frequencies converges faster and has low variance. It is

Table 2: Log likelihood estimates of EvoVGM models using validation alignments of five sequences with a length of 5000 bp. JC69, K80 and GTR substitution models were used to simulate training and validation alignments. The estimates are computed and averaged from fitting and running the models ten times.

	JC69		K80		GTR	
	MEAN	STD	MEAN	STD	MEAN	STD
ACTUAL	-17249.830		-17024.340		-15818.739	
EvoVGM_JC69	-17209.913	142.128	-17287.278	185.441	-16491.810	125.664
EvoVGM_K80	-17203.100	151.758	-17007.724	133.294	-16495.530	175.024
EvoVGM_GTR	-17204.540	121.459	-17014.296	151.457	-15540.730	126.673

more accurate when the model is trained with larger sequence alignments.

We present in Table 1 the performance of trained EvoVGM_GTR models in approximating the evolutionary parameters from new multiple alignments, which were simulated with the same set of evolutionary parameters used in the training step. We noted that the estimated branch lengths are close and strongly correlated to their actual values for all datasets except the smallest. Also, we noted that EvoVGM_GTR estimates better the relative frequencies than the substitution rates in the small datasets. However, as the datasets get larger, the approximation of the substitution rates gets better.

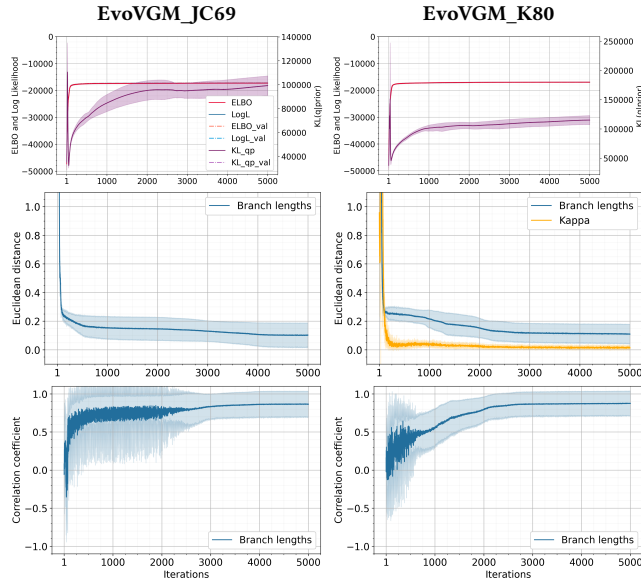


Figure 4: Convergence and performance of EvoVGM_JC69 and EvoVGM_K80 models using alignments of five sequences with a length of 5000 bp. JC69 and K80 substitution models were used to simulate training and validation alignments. The results are computed and averaged from fitting and running the models ten times.

Additionally, we evaluated the convergence and the accuracy of EvoVGM_JC69 and EvoVGM_K80, two variants of the EvoVGM

model implementing JC69 and K80 substitution models, respectively. Each model was fitted ten times with different weight initialization on the exact sequence alignment. Figure 4 and Table 2 show the results of the models trained and evaluated with alignments of five sequences with a length of 5000 bp. All models converge to values closer to or higher than the actual log likelihood of the data, which is calculated with equation 3. Moreover, they were able to approximate the branch lengths even when trained with datasets simulated with a different substitution model (Tables A.1 and A.2).

4.3 Estimating Evolutionary Parameters on Real Alignment

Finally, we analyzed a real dataset to assess the robustness of the estimation provided by the variational generative model EvoVGM. The dataset was recovered from [23], and it consists of six sequences of Gene S of coronaviruses. We used the NGPhylogeny.fr platform [18] to build a multiple sequence alignment with MAFFT 7.407 [12]. The alignment has a length of 3688 bp after cleaning the sequences from gaps using Gblocks 0.91.1 [26].

We applied EvoVGM_GTR model to the dataset using the same configuration of the variational encoders and the hyper-parameters defined previously. We set gamma priors on branch lengths with a prior expectation of 0.01. We found that using α_{KL} with a value of 0.1 gives better estimations but with higher variance. We trained EvoVGM_GTR over 5000 iterations and replicated it ten times. Furthermore, we compared the estimations of EvoVGM_GTR model with those of MrBayes 3.2.7, a Bayesian phylogenetic inference program [9]. We ran MrBayes with four chains and two runs for one million iterations and sampling every 500 iterations.

In Figure 5, we show the evolution of the Euclidean distance and the correlation coefficient between the estimations of EvoVGM_GTR and those of MrBayes during the training of EvoVGM_GTR. The estimations of the branch lengths differ from those of MrBayes with a distance lower than 0.2 but with high variance. Furthermore, EvoVGM_GTR managed to estimate the substitution rates and the relative frequencies with low-variance values closer to the estimations of MrBayes.

5 CONCLUSION

In this work, we show that a deep variational Bayesian generative method could constitute a feasible option to approximate the true parameters of an evolutionary model and generate the associated sequence alignment. The implementation of this method,

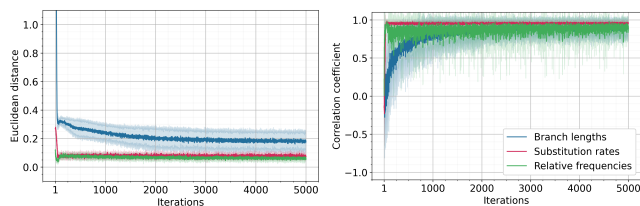


Figure 5: Comparison of EvoVGM_GTR with MrBayes. The results are averaged from fitting EvoVGM_GTR ten times. They are reported in terms of Euclidean distance and correlation coefficient of EvoVGM_GTR and MrBayes estimated evolutionary parameters.

EvoVGM, estimates the branch lengths, the ancestral states, and the substitution model parameters from a multiple sequence alignment. We assessed its consistency and effectiveness using sequence alignments simulated with different sizes. In general, the EvoVGM model needs a few thousand iterations to converge. It tends to be accurate with low variance in estimating the evolutionary parameters using fine-tuned hyper-parameters. Moreover, it provides an effective way of estimating the parameters for different substitution models such as JC69, K80, and GTR. The generalization to other models like HKY is also straightforward. For future work, many extensions could be explored to improve the EvoVGM model, such as considering a prior tree topology, investigating the influence of the priors on inference, and allowing parameter heterogeneity across sites and lineages.

ACKNOWLEDGMENTS

We would like to thank Golrokh Vitae, Hayda Almeida, Maia Kaplan, Dylan Lebatteux, Mathieu Blanchette and Vladimir Makarenkov for their helpful discussions.

This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada. It has also been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec - Nature et technologies (FRQNT), Génome Québec and Genome Canada for the grants to ABD. AMR received NSERC and FRQNT scholarships during the development of this work.

REFERENCES

- [1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.
- [2] Tung Dang and Hirohisa Kishino. 2019. Stochastic Variational Inference for Bayesian Phylogenetics: A Case of CAT Model. *Molecular Biology and Evolution* 36, 4 (apr 2019), 825–833.
- [3] Abdoulaye Banire Diallo, Vladimir Makarenkov, and Mathieu Blanchette. 2007. Exact and heuristic algorithms for the indel maximum likelihood problem. *Journal of Computational Biology* 14, 4 (2007), 446–461.
- [4] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, et al. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544, 7650 (2017), 309–315.
- [5] Nuno R Faria, Andrew Rambaut, Marc A Suchard, Guy Baele, Trevor Bedford, Melissa J Ward, Andrew J Tatem, João D Sousa, Nimalan Arinaminpathy, Jacques Pépin, et al. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *science* 346, 6205 (2014), 56–61.
- [6] Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* 17, 6 (1981), 368–376.
- [7] Mathieu Fourment and Aaron E. Darling. 2019. Evaluating probabilistic programming and fast variational Bayesian inference in phylogenetics. *PeerJ* 7, 12 (dec 2019), e8272.
- [8] John P. Huelsenbeck, Bret Larget, Richard E. Miller, and Fredrik Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* 51, 5 (2002), 673–688.
- [9] John P. Huelsenbeck and Fredrik Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 8 (2001), 754–755.
- [10] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37, 2 (1999), 183–233.
- [11] Thomas H Jukes and Charles R Cantor. 1969. Evolution of protein molecules. In *Mammalian protein metabolism*, H. H. Munro (Ed.). Vol. III. Academic Press, New York, 21–132.
- [12] Kazutaka Katoh and Daron M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30, 4 (01 2013), 772–780.
- [13] Andrew D Kern and David Haussler. 2010. A population genetic hidden Markov model for detecting genomic regions under selection. *Molecular biology and evolution* 27, 7 (2010), 1673–1685.
- [14] Motoo Kimura. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* 16, 2 (1980), 111–120.
- [15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
- [16] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations*.
- [17] Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme. 2009. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.
- [18] Frédéric Lemoine, Damien Correia, Vincent Lefort, Olivia Doppelt-Azeroual, Fabien Mareuil, Sarah Cohen-Boulakia, and Olivier Gascuel. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Research* 47, W1 (04 2019), W260–W265.
- [19] Dongjoon Lim and Mathieu Blanchette. 2020. EvoLSTM: context-dependent models of sequence evolution using a sequence-to-sequence LSTM. *Bioinformatics* 36, Supplement_1 (jul 2020), i353–i361.
- [20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035.
- [21] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *31st International Conference on Machine Learning, ICLR 2014* 4 (2014), 3057–3070.
- [22] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* 15, 10 (2018), 816–822.
- [23] Stéphane Samson, Étienne Lord, and Vladimir Makarenkov. 2022. SimPlot++: a Python application for representing sequence similarity and detecting recombination. *Bioinformatics* (04 2022), btac287.
- [24] David Sankoff and Mathieu Blanchette. 1999. Probability models for genome rearrangement and linear invariants for phylogenetic inference. In *Proceedings of the third annual international conference on Computational molecular biology*, 302–309.
- [25] Stephanie J. Spielman and Claus O. Wilke. 2015. Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE* 10, 9 (09 2015), 1–7.
- [26] Gerard Talavera and Jose Castresana. 2007. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology* 56, 4 (08 2007), 564–577.
- [27] Simon Tavaré et al. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* 17, 2 (1986), 57–86.
- [28] Eli N Weinstein and Debora Marks. 2021. A structured observation distribution for generative biological sequence prediction and forecasting. In *International Conference on Machine Learning*. PMLR, 11068–11079.
- [29] Chris Whidden and Frederick A Matsen IV. 2015. Quantifying MCMC exploration of phylogenetic tree space. *Systematic biology* 64, 3 (2015), 472–491.
- [30] Ziheng Yang. 1994. Estimating the pattern of nucleotide substitution. *Journal of molecular evolution* 39, 1 (1994), 105–111.

- [31] Ziheng Yang. 2014. *Molecular evolution: a statistical approach*. Oxford University Press.
- [32] Ziheng Yang and Bruce Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular biology and evolution* 14, 7 (1997), 717–724.
- [33] Ke Yuan, Thomas Sakoparnig, Florian Markowitz, and Niko Beerenwinkel. 2015. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology* 16, 1 (2015), 1–16.
- [34] Cheng Zhang. 2020. Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18760–18771.
- [35] Cheng Zhang and Frederick A. Matsen IV. 2019. Variational Bayesian Phylogenetic Inference. In *International Conference on Learning Representations*.

A APPENDIX

A.1 Development of the ELBO $\mathcal{L}(\phi, X)$

$$\begin{aligned}
\log p(X) &= \sum_{n=1}^N \log p(x_n) \\
&= \sum_{n=1}^N \mathbb{E}_{q_\phi(a_n, t, \psi | x_n)} \left[\log \frac{p(x_n, a_n, t, \psi)}{p(a_n, t, \psi | x_n)} \right] \\
&= \sum_{n=1}^N \mathbb{E}_{q_\phi(a_n, t, \psi | x_n)} \left[\log \frac{p(x_n, a_n, t, \psi)}{q_\phi(a_n, t, \psi | x_n)} \frac{q_\phi(a_n, t, \psi | x_n)}{p(a_n, t, \psi | x_n)} \right] \\
&= \sum_{n=1}^N \mathbb{E}_{q_\phi} \left[\log \frac{p(x_n, a_n, t, \psi)}{q_\phi(a_n, t, \psi | x_n)} \right] + \mathbb{E}_{q_\phi} \left[\log \frac{q_\phi(a_n, t, \psi | x_n)}{p(a_n, t, \psi | x_n)} \right] \\
&= \underbrace{\sum_{n=1}^N \mathcal{L}_n(\phi, x_n)}_{\geq \mathcal{L}(\phi, X)} + \sum_{n=1}^N \text{KL} \left(q_\phi(a_n, t, \psi | x_n) \parallel p(a_n, t, \psi | x_n) \right) \\
&\geq \mathcal{L}(\phi, X).
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(\phi, X) &= \sum_{n=1}^N \mathcal{L}_n(\phi, x_n) \\
&= \sum_{n=1}^N \mathbb{E}_{q_\phi} \left[\log \frac{p(x_n, a_n, t, \psi)}{q_\phi(a_n, t, \psi | x_n)} \right] \\
&= \sum_{n=1}^N \mathbb{E}_{q_\phi} \left[\log p(x_n | a_n, t, \psi) + \log p(a) + \log p(t) + \log p(\psi) \right. \\
&\quad \left. - \log q_{\phi_a}(a_n | x_n) - \log q_{\phi_t}(t) - \log q_{\phi_\psi}(\psi) \right] \\
&= -N \left(\mathbb{E}_{q_\phi} \left[\log p(\psi) - \log q_{\phi_\psi}(\psi) \right] + \mathbb{E}_{q_\phi} \left[\log p(t) - \log q_{\phi_t}(t) \right] \right) \\
&\quad + \sum_{n=1}^N \mathbb{E}_{q_\phi} \left[\log p(x_n | a_n, t, \psi) \right] + \mathbb{E}_{q_\phi} \left[\log p(a) - \log q_{\phi_a}(a_n | x_n) \right] \\
&= -N \left(\text{KL}(q_{\phi_\psi}(\psi) \parallel p(\psi)) + \text{KL}(q_{\phi_t}(t) \parallel p(t)) \right) \\
&\quad + \sum_{n=1}^N \mathbb{E}_{q_\phi} \left[\log p(x_n | a_n, t, \psi) \right] - \text{KL}(q_{\phi_a}(a_n | x_n) \parallel p(a)) \\
&= -N \left(\text{KL}(q_{\phi_\psi}(\psi) \parallel p(\psi)) + \sum_{m=1}^M \text{KL}(q_{\phi_t}(t^m) \parallel p(t)) \right) \\
&\quad + \sum_{n=1}^N \left(\frac{1}{L} \sum_{l=1}^L \sum_{m=1}^M \log p(x_n^m | a_n^l, t^{m,l}, \psi^l) \right) - \text{KL}(q_{\phi_a}(a_n | x_n) \parallel p(a)).
\end{aligned}$$

A.2 Supplemental Results

Table A.1: Distance and correlation between actual and estimated branch lengths by EvoVGM_JC69.

$N \rightarrow$	100			1000			5000		
M	DIST	CORR	PVAL	DIST	CORR	PVAL	DIST	CORR	PVAL
3	0.129	0.969	0.160	0.069	0.982	0.121	0.143	0.982	0.120
4	0.166	0.938	0.062	0.065	0.997	0.003	0.079	0.997	0.003
5	0.179	0.841	0.074	0.096	0.993	0.001	0.076	0.990	0.001

Table A.2: Distance and correlation between actual and estimated branch lengths by EvoVGM_K80.

$N \rightarrow$	100			1000			5000		
M	DIST	CORR	PVAL	DIST	CORR	PVAL	DIST	CORR	PVAL
3	0.133	0.948	0.207	0.171	0.975	0.144	0.074	0.975	0.142
4	0.184	0.855	0.145	0.093	0.996	0.004	0.049	0.992	0.008
5	0.180	0.835	0.078	0.062	0.990	0.001	0.073	0.999	0.000

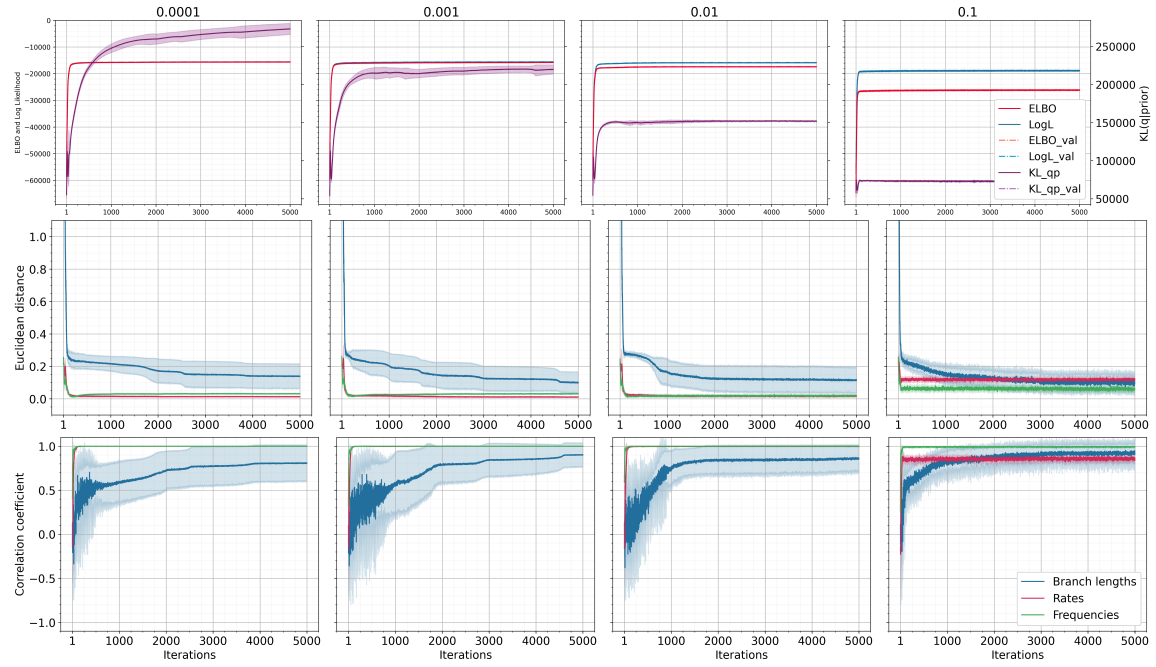


Figure A.1: Convergence and performance of EvoVGM_GTR model for multiple settings of α_{KL} . The GTR substitution model was used to simulate training and validation alignments of five sequences with a length of 5000 bp. The estimates are computed and averaged from fitting and running the model ten times.

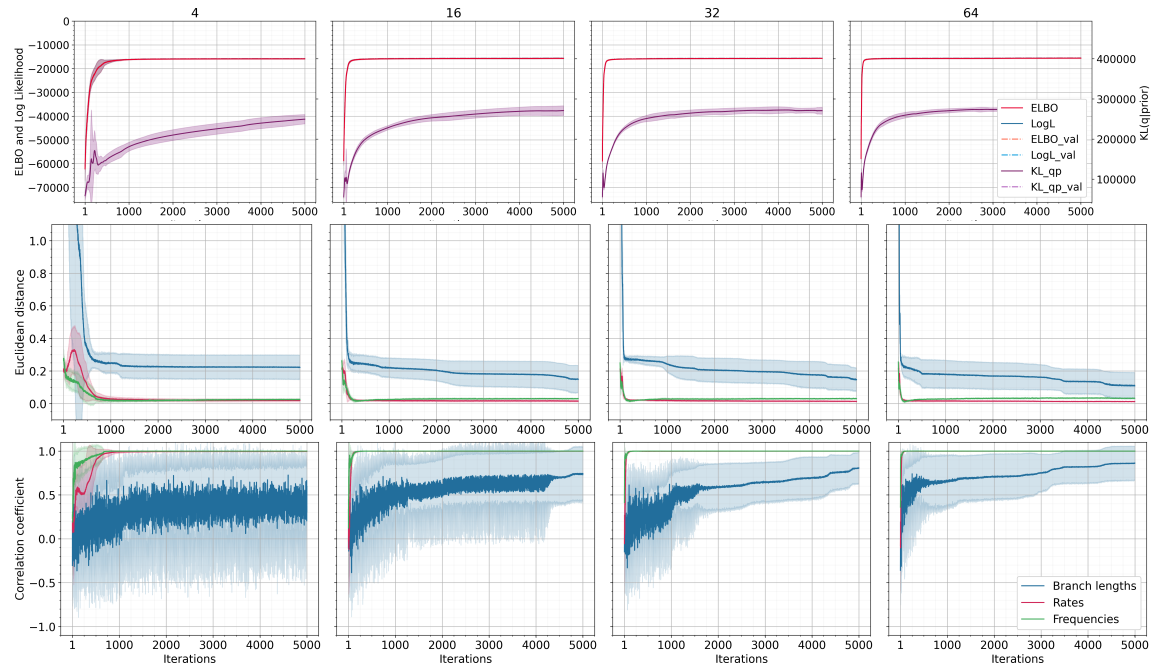


Figure A.2: Convergence and performance of EvoVGM_GTR model for multiple settings of the hidden size of the neural networks of the variational encoders. The GTR substitution model was used to simulate training and validation alignments of five sequences with a length of 5000 bp. The estimates are computed and averaged from fitting and running the model ten times.

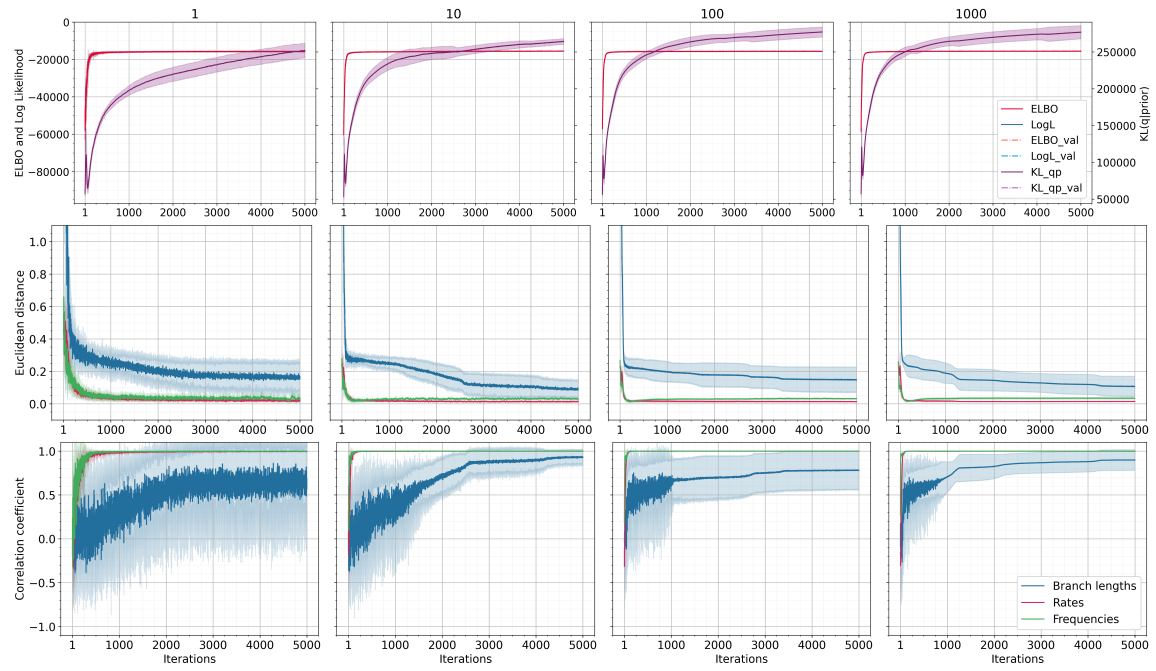


Figure A.3: Convergence and performance of EvoVGM_GTR model for multiple settings of the sample size. The GTR substitution model was used to simulate training and validation alignments of five sequences with a length of 5000 bp. The estimates are computed and averaged from fitting and running the model ten times.

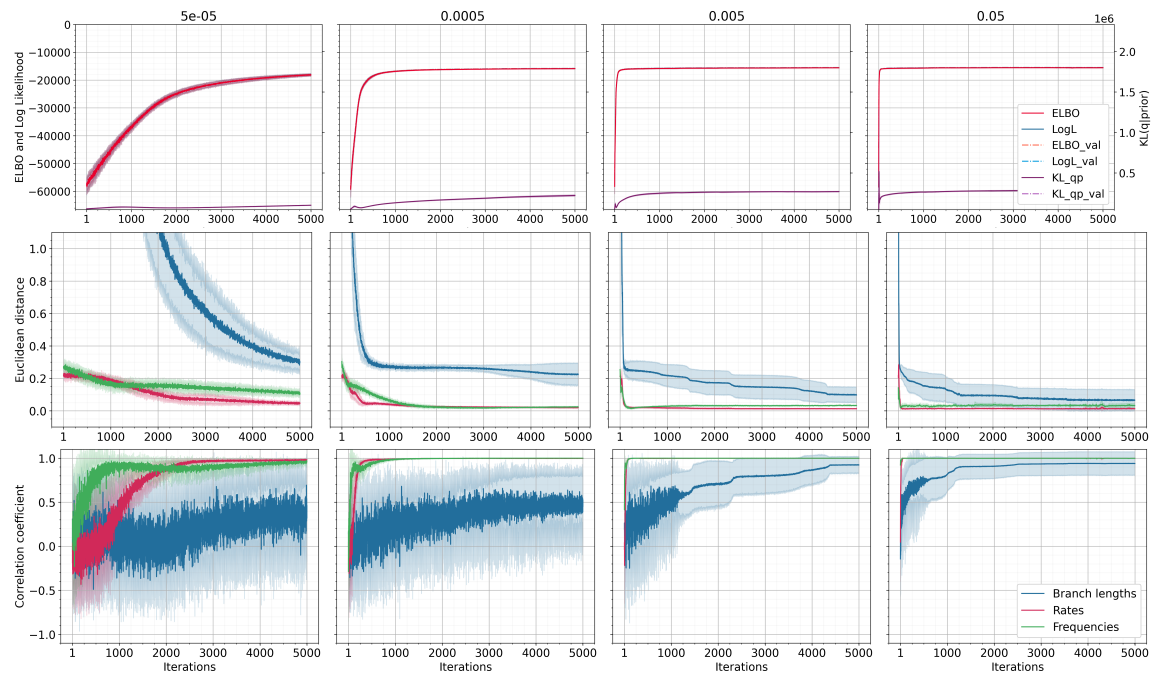


Figure A.4: Convergence and performance of EvoVGM_GTR model for multiple settings of the learning rate. The GTR substitution model was used to simulate training and validation alignments of five sequences with a length of 5000 bp. The estimates are computed and averaged from fitting and running the model ten times.