Stochastic Variational Inference for Bayesian Phylogenetics: A Case of CAT Model

Tung Dang,[*,1] *and Hirohisa Kishino*[1]
[1] Laboratory of Biometrics and Bioinformatics, University of Tokyo, Tokyo, Japan
[*]Corresponding author: dangthanhtung91@vn-bml.com

Abstract

The pattern of molecular evolution varies among gene sites and genes in a genome. By taking into account the complex heterogeneity of evolutionary processes among sites in a genome, Bayesian infinite mixture models of genomic evolution enable robust phylogenetic inference. With large modern data sets, however, the computational burden of Markov chain Monte Carlo sampling techniques becomes prohibitive. Here, we have developed a variational Bayesian procedure to speed up the widely used PhyloBayes MPI program, which deals with the heterogeneity of amino acid propensity. Rather than sampling from the posterior distribution, the procedure approximates the (unknown) posterior distribution using a manageable distribution called the variational distribution. The parameters in the variational distribution are estimated by minimizing Kullback-Leibler divergence. To examine performance, we analyzed three large data sets consisting of mitochondrial, plastid-encoded, and nuclear proteins. Our variational method accurately approximated the Bayesian phylogenetic tree, mixture proportions, and the amino acid propensity of each component of the mixture while using orders of magnitude less computational time.

1 Introduction

Understanding the evolutionary variation of phenotypic characters and testing hypotheses about the underlying mechanism are some of the main concerns of evolutionary biology. Because this variation needs be interpreted as an evolutionary history, accurately inferring the phylogenetic tree is important. Otherwise, the uncertainty of phylogenetic inference must be taken into account to obtain an unbiased picture of evolutionary variation.

The increasing amount of available genomic data enables reliable inference of phylogenetic trees. Because molecular evolution is largely driven by nearly neutral or slightly deleterious mutations Ohta (1973), this process is less prone to convergent evolution compared with the evolution of phenotypic traits. The pattern of molecular evolution is statistically formulated by Markov processes. The pattern and rate of molecular evolution is complex, however, depending on various factors affecting mutation rates and functional constraints. To model protein evolution, Thorne, Goldman, and Jones (1996) introduced the concept of hidden states of secondary structure to describe sites of heterogeneity Goldman *et al.* (1996); Thorne *et al.* (1996); Jones *et al.* (1996). Koshi and Goldstein (1998) developed a model of physico-chemical properties of amino acids, while Halpern and Bruno (1998) introduced a more advanced model with position-specific amino acid frequencies.

Equilibrium amino acid frequencies, which reflect structural and functional constraints, vary among sites within and among proteins. Inter-species comparative genomics approaches can analyze a huge number of alignment columns, but the number of taxa is often insufficient to estimate individual position-specific amino acid frequencies. To achieve a balance between variance and bias, Lartillot and Philippe (2004) proposed a Bayesian non-parametric approach based on a countable infinite mixture model, referred to as the CAT model. This model specifies K of distinct processes (or classes), each characterized by a particular set of equilibrium frequencies, and sites are distributed according to a mixture of these K distinct processes. By

proposing a truncated stick-breaking representation of the Dirichlet process prior on the space of equilibrium frequencies (Ferguson 1973; Green and Richardson 2001; Ishwaran and James 2001), the total number of classes can be treated as free variables of the model. A hybrid framework between Gibbs-sampling and Metropolis-Hastings algorithm have been developed to estimate all parameters of the model Papaspiliopoulos and Roberts (2008).

Existing approaches cannot take full advantage of the CAT model (Lartillot and Philippe 2004; Lartillot 2006), because the computational burden is prohibitive for inference based on large data sets. Even well-designed sampling schemes need to generate a large number of posterior samples through the entire data set to resolve convergence, and their convergence can be difficult to diagnose. To provide faster estimation, Lartillot *et al.* (2013) developed a message passing interface (MPI) for parallelization of the PhyloBayes MPI program. By implementing Markov chain Monte Carlo (MCMC) samplers in a parallel environment, PhyloBayes MPI allows for faster phylogenetic reconstruction under complex mixture models.

Here, we propose an alternative approach, a variational inference method (Jordan *et al.* 1999; Bishop 2006; Blei *et al.* 2006; Hoffman *et al.* 2013). The basic idea of variational inference is the formulation of the estimation of marginal or conditional probabilities as an optimization problem rather than sampling-based inference. Variational methods, originally used in statistical physics to approximate intractable integrals, have been successfully used in wide variety of applications related to complex networks (Gopalan and Blei 2013) and population genetics (Gopalan *et al.* 2016; Raj *et al.* 2014). In this article, we demonstrate that our algorithms are considerably faster than PhyloBayes MPI while achieving comparable accuracies.

2 New Approaches

The CAT model formulates the substitutional heterogeneity across sites of protein sequences as a mixture of different equilibrium amino acid frequencies, called profiles. By introducing a Dirichlet process prior on these profiles, the number of categories, the profile of each category, and the resultant phylogenetic tree are estimated from the data in a Bayesian framework. The standard Markov chain Monte Carlo (MCMC) approach facilitates parameter estimation of this parameter-rich model and enables robust inference of phylogenetic trees while allowing for the complexity of protein evolution.

The rapid growth of genomic databases theoretically enables accurate classification of amino acid sites in protein sequences, but the Monte Carlo integration becomes computationally more challenging. To allow the CAT model to extract the maximum amount of relevant information from the data, we have developed a variational Bayesian procedure. The core of the variational framework is a mean-field approximation of the posterior distribution. We approximate the posterior distribution with a mean field representation of the variational distribution, which is much easier to work with computationally. In this approximation, the parameters and hidden variables are assumed to be independent of one another. The parameters of the variational distribution are obtained by minimizing the Kullback-Leibler (KL) divergence between the true conditional distributions of the hidden variables given the observations and their variational distributions. Inference becomes a single optimization problem that gives us approximate analytical forms for the posterior distributions over unknown variables of the CAT model as well as an approximate estimate of the intractable marginal likelihood. To deal with the uncertainty of tree topologies, we have preserved the Gibbs sampling algorithm of tree topologies (Lartillot *et al.* 2013).

Table 1: **Run times of variational inference and MCMC algorithms on real data**

Data set	Taxa	Sites	States	MCMC	VI
Data Set A	13	6,622	20	4.72 days	0.81 day
Data Set B	28	10,137	20	10.61 days	2.36 days
Data Set C	66	38,330	20	28.35 days	5.67 days

Both variational inference and MCMC algorithms were run in a parallel environment. The properties of the parallel version were evaluated on a personal computer (Intel Core i7-6700 CPU 3.40GHz, 8 cores, 2 threads per core, 4 cores per socket, 16 Gb RAM), under Linux Mint 17.3 Rosa.

3 Results

3.1 Runtime Performance

To compare the performance of our version of variational inference with that of the MCMC algorithm of PhyloBayes MPI, we estimated the CAT model with both algorithms using real data sets. This portion of the study was carried out using three real data sets, the largest consisting of 38,330 amino acid positions from 66 species. The goals of this data analysis were to demonstrate the numerical feasibility of our implementations and to ascertain the accuracy of our variational inference approach. In our comparisons, all algorithms were timed under equivalent computational conditions. Because of the intensive nature of the estimations, further computational experiments will be required to test the performance of variational inference on much more massive data sets.

First, we explored whether our new approximation approach could significantly reduce the computational burden required to estimate all parameters of the CAT model. We focused our analysis on the three real data sets described in detailed Data Sets5.6.

Table 1 illustrates the computational time required for estimation of all parameters in the CAT-Poisson model when optimized using variational inference compared with sampling under the MCMC algorithm across three data sets. These data sets contained drastically different numbers of taxa and sites. For example, the number of taxa and sites in data set C were approximately three times larger than those in data set B. The time complexity of each of the above algorithms was found to increase regularly with the number of genes, species and total aligned amino acid positions. Run times were significantly reduced in the variational inference framework compared with those in the MCMC approach. While our procedure uses the variational inference procedure to estimate parameters of the evolutionary process, we note that we have retained the algorithm for Gibbs sampling of tree topologies. If this partial MCMC algorithm can also be replaced by some other optimization, the computation burden will be greatly reduced.

3.2 Accuracy of Estimated Topologies, Tree Lengths, and Profiles

The tree topology and branch lengths estimated by variational inference were almost the same as those obtained by the MCMC algorithm (Figure 1).

By introducing a Dirichlet process prior, the CAT model provides a posterior distribution of K, the number of separate categories, and the size of each category. The PhyloBayes MPI program, which is based on a hybrid strategy between Gibbs sampling and Metropolis-Hastings algorithm, first proposes allocation variables and stationary probabilities at all other sites. These site to category reallocation proposals, which are driven by the posterior weights of the mixture and profiles associated with each component of the mixture, are performed by Gibbs sampling. Metropolis-Hastings algorithms are then used to consider the classes for sites. This

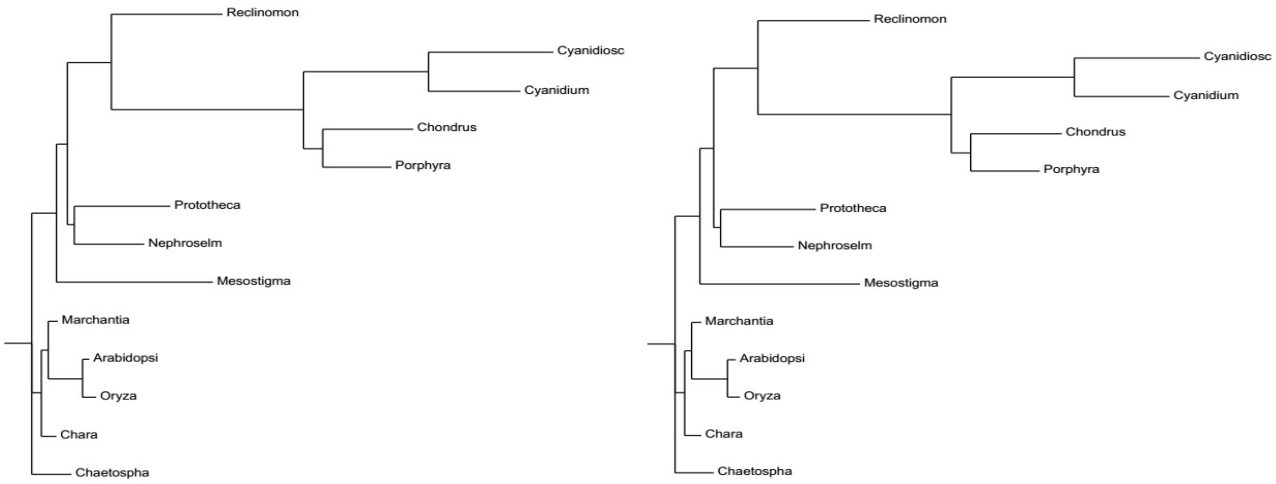

Figure 1: Posterior consensus trees between MCMC (left) and variational inference (right) approaches obtained using a mitochondrial data set (13 taxa and 6,622 amino acid positions (Rodríguez-Ezpeleta *et al.* 2006))

Table 2: **Classes estimated by variational inference and MCMC in data set A**

No. Class (Number of Sites)			
MCMC	VI	MCMC	VI
9 (524)	59 (527)	12 (256)	48 (246)
4 (481)	4 (480)	8 (235)	40 (240)
23 (457)	92 (454)	11 (226)	32 (220)
21 (403)	82 (400)	19 (197)	26 (188)
5 (328)	5 (326)	35 (161)	63 (157)
16 (284)	33 (290)	1 (148)	1 (145)
18 (273)	39 (276)	22 (140)	35 (137)
15 (265)	28 (275)	2 (78)	2 (76)

Top-ranked estimated classes are listed along with the number of sites distributed in each class. The results are for real data set A, with the number of sites calculated by counting sites allocated to each class.

strategy guarantees that the samplers leave the posterior distribution invariant. Our approach, variational inference, proposes variational distributions for allocation variables and weights of the mixture and profiles. The choice among alternative allocations of sites to categories is driven by updating parameters of these variational distributions and computing the expected values of these variables under variational distribution.

Table 2 compares some major categories estimated by MCMC and variational inference. The size of each category was approximated by the number of sites assigned to that class. The number of distinct categories was estimated for data set A representing 6,622 amino acid positions. As can be seen in the table, variational inference accurately approximated posterior mean sizes of these categories, their profiles were accurately estimated as well (Figure 2).

Taken together, these results demonstrate that the estimation time required by the variational inference framework compares favorably with that used by sampling algorithms such as MCMC, while a sufficient level of accuracy under the CAT model is still guaranteed.

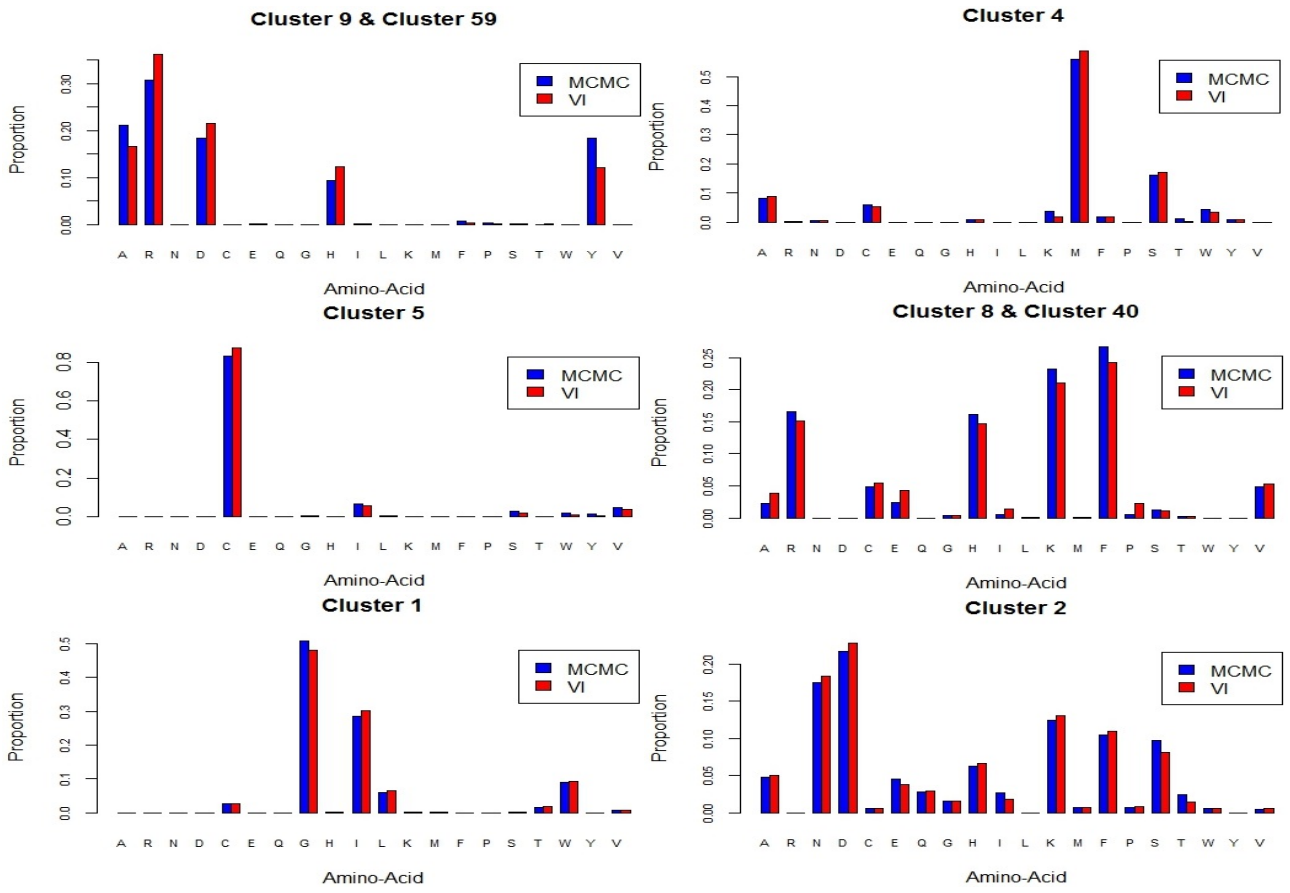

Figure 2: Equilibrium frequency profiles of categories determined by MCMC and variational inference algorithms based on a mitochondrial data set (13 taxa and 6,622 amino acid positions (Rodríguez-Ezpeleta *et al.* 2006)). In the case of the first few categories, the two algorithms performed similarly. A detailed comparison of the remaining categories between MCMC and variational inference can be found in Table 2.

4 Discussion

We have developed a new framework for estimating all parameters of the CAT model, namely, stochastic variational inference, that can considerably improve runtime performance as well as significantly reduce the computational burden. In contrast to existing approaches designed for the same purpose that rely on simulation framework, such as Gibbs-sampling and Metropolis-Hastings algorithms (Lartillot and Philippe 2004; Lartillot 2006; Lartillot *et al.* 2013), stochastic variational inference recasts the problem of inference as an optimization problem, thus allowing us to design powerful tools for convex optimization. In this way, our approach proposes a feasible family of variational distributions and then selects the family member closest to the true intractable posterior distribution of interest by optimizing Kullback-Leibler (KL) divergence.

We have demonstrated through analysis of actual data sets that our method accurately approximates the posterior distribution of the CAT model with improved speed. This substantial runtime enhancement with no loss of accuracy allows our method to be applied to the large data sets that are steadily becoming the norm in phylogenetic and biological evolutionary studies. Finally, our results were obtained on a modest computing platform. The implementation of a variational inference version of PhyloBayes MPI to exploit advanced computing architectures holds the promise of analyzing even larger data sets than the examples in our paper.

Bayesian models of sequence evolution allow substitutional heterogeneity across protein sequence sites to be taken into account. In particular, the CAT model treats the number of substitutional categories as a free parameter and is able to uncover a level of heterogeneity

much higher than that assumed by other mixture models. Under the variational inference approach, all of these special features of the CAT model are guaranteed. Given the increasing size of studied data sets, ensuring statistical algorithms scale to a large number of species with massive numbers of nucleotide positions is critical. We have shown that such analyses are time consuming when undertaken with MCMC algorithms, which perform a large number of multiple simulated iterations over the entire data set. With their more efficient optimization framework, stochastic variational inference algorithms overcome this limitation without compromising all of the principles and statistical assumptions behind the model. For improved estimations of site-heterogeneous Bayesian mixture models with the massive data set, we recommend implementation of a variational inference version of PhyloBayes MPI.

5 Materials and Methods

5.1 CAT Model

We use an infinite mixture model that describes site heterogeneity with respect to the substitution process. This model is similar to one proposed by Lartillot and Philippe (2004), but, instead of sampling-based inference, we have developed a new approach that allows for efficient inference of ancestral sequences. Our model, which does not assume that all sites of a protein evolve under the same substitution process, is characterized by a 20x20 substitution matrix. In addition, the model does not assume a fixed number of distinct substitution processes (or classes) and respective amino-acid profiles; instead, these are treated as free variables of the model. Poisson (or F81) Felsenstein (1981) Markov processes are considered to apply to all substitution processes along the branches of a tree (Lartillot and Philippe 2004; Lartillot 2006). Each Markov process is characterized by a rate matrix $Q = [Q_{ab}]$, which can be expressed in terms of a vector of stationary probabilities, or equilibrium frequencies $\pi_a, 1 \leq a \leq 20$ such that $\sum_{a=1}^{20} \pi_a = 1$ and a set of relative rates, or exchangeability parameters, $(\rho_{ab}), 1 \leq a, b \leq 20$. We determine the size of the segments adaptively as described below. This approach allows us to work in the framework of a Bayesian mixture model with parameters representing the mixture of distinct classes, the rates at each site. and branch lengths.

Given an amino-acid database including N aligned positions (columns) and P taxa, we label the data matrix D_{ip} as simply the possible states of the process operating at site i for i = 1, ..., N at the leaf indexed by p $(1 < p < P)$. We consider $l_j\,(1 < j < 2P - 3)$; $r_i\,(1 \leq i \leq N)$ to be random variables that denote the branch j and the relative rate of substitution at each site i. Formally, the CAT-Poisson model assumes that (i) a gamma distribution of shape 1 and scale $\beta > 0$ is the prior distribution of branch lengths, (ii) a gamma distribution of shape α and scale α is the prior distribution of rates, and (iii) the prior distribution on profile π is a flat Dirichlet distribution. Furthermore, Lartillot *et al.* (2013) has developed a Dirichlet process mixture model formulated in terms of a stick-breaking construction over the equilibrium frequency profile to generate an infinite number of mixtures of Poisson processes for describing sites, with each mixture characterized by its own substitution matrix $\left\{Q^k\right\}, k = 1, ..., \infty$ and only the stationary probabilities $\left(\pi_a^k\right), a \in [1, ..., 20], k \in [1, ..., \infty]$ differing. By proposing a new random variable V_k which is the unit length of the k^{th} stick, the stick-breaking representation allows the construction of an infinite mixture structure. Moreover, each site i in an amino-acid sequence belongs to a category k that is specified by the allocation variable $z_i \in [1, ..., \infty]$. The vector $z = (z_i)$ where $i \in [1, ..., N]$, is called the allocation vector. The allocations $z = (z_i)$ are drawn i.i.d from a multinomial of the infinite vector of mixing proportions, namely, $\varphi = (\varphi_k), k \in [1, ..., \infty]$. In addition, we use a data augmentation algorithm, that is proposed (Nielsen 2002), to obtain the substitution mapping in the case of Poisson processes. The

substitution mapping is described by the formula $\Xi_{ij} = \left(n_{ij}, \left(\sigma_{ij}^h\right)_{h=1,\ldots,n_{ij}-1}\right)$, n_{ij} denotes the number of substitutions on branch j and at site i, $\left(\sigma_{ij}^h\right)_{h=1,\ldots,n_{ij}-1}$ denotes the successive states of the process and random variable w_a^k is the total number of substitutions to state a at sites which are assigned in cluster k, plus one if a is the state at the root of the tree. This algorithm is used to simulate the mutational history for a single site. The probability distribution of n_{ij} is defined by the Poisson distribution of rate parameter $r_i l_j$ and $\left(\sigma_{ij}^h\right)_{h=1,\ldots,n_{ij}-1}$ is drawn from $\left(\pi_a^k\right), a \in [1, ..., 20], k \in [1, ..., \infty]$.

Given a data set of amino-acid sequences, Markov chain Monte Carlo (MCMC) sampling methods have been proposed to approximate full parameters of this model (Lartillot and Philippe 2004; Lartillot 2006). A parallel computing version has been developed to speed up the estimation process, thus allowing faster inference the phylogenetic reconstruction under a Dirichlet Mixture Process (Lartillot *et al.* 2013). Basically, however, MCMC methods, even parallel MCMC, solve this problem based on sampling schemes from a Markov chain whose stationary distribution is the posterior of interest and by updating an estimate of the model parameters. When a database becomes too large for memory or iterative computation, these approaches significantly increase the time complexity of inference.

5.2 Variational Inference

Variational inference is a class of methods that reformulate the problem of approximating the posterior inference for complex probabilistic models as an optimization problem. The central purpose of the variational inference algorithm is to approximate the true intractable posterior distribution $p(\Phi, \Xi|D)$, $\Phi = \{V, z, \pi, l, r\}$ by finding an element of a tractable family of probability distributions $q(\Phi, \Xi|\Theta)$, called the *variational distribution*. These distributions are parameterized by free parameters, called *variational parameters* Θ. Variational inference fits these parameters to find a distribution close to the true intractable posterior distribution of interest. The distance on probability space for a pair of probability distribution $q(\Phi, \Xi|\Theta)$ and $p(\Phi, \Xi|D)$ is measured with Kullback-Leibler (KL) divergence:

$$\begin{aligned} &KL\left[q(\Phi, \Xi|\Theta)|p(\Phi, \Xi|D)\right] \\ &= \mathrm{E}_q\left[\log\left\{q(\Phi, \Xi|\Theta)\right\}\right] - \mathrm{E}_q\left[\log\left\{p\left(\Phi, \Xi|D\right)\right\}\right] \\ &= \mathrm{E}_q\left[\log\left\{q(\Phi, \Xi|\Theta)\right\}\right] - \mathrm{E}_q\left[\log\left\{p\left(D, \Phi, \Xi\right)\right\}\right] \\ &+ \log p\left(D\right). \end{aligned} \tag{1}$$

The term $\log p\left(D\right)$ in equation (1), which is the cause of computational difficulty in Bayesian analysis, can be treated as a constant to estimate the variational distribution that is closest to the posterior distribution:

$$q^*(\Phi, \Xi|\Theta) = argmin KL\left[q(\Phi, \Xi|\Theta)|p(\Phi, \Xi|D)\right].$$

By adopting the compromised target function $KL\left[q(\Phi, \Xi|\Theta)|p(\Phi, \Xi|D)\right]$, the variational inference maximizes the computational feasible target function:

$$\begin{aligned} &\mathcal{L}\left[q(\Phi, \Xi|\Theta)\right] \\ &= \mathrm{E}_q\left[\log\left\{p\left(D, \Phi, \Xi\right)\right\}\right] - \mathrm{E}_q\left[\log\left\{q(\Phi, \Xi|\Theta)\right\}\right]. \end{aligned} \tag{2}$$

Because

$$\log p\left(D\right) = \mathcal{L}\left[q(\Phi, \Xi|\Theta)\right] + KL\left[q(\Phi, \Xi|\Theta)|p(\Phi, \Xi|D)\right],$$

the equation (2) is called Evidence Lower BOund (ELBO (Jordan *et al.* (1999))) It should be noted that the value of the target function cannot be used for comparison between different models of variational functions. Currently, the standard model checking process is to compare the important aspects of $q^*(\Phi, \Xi|\Theta)$ with those of MCMC runs by example data.

5.3 Two Illustrative Examples

5.3.1 Variational Inference of Bayesian Ridge Regression

We consider the linear regression model. Given a data set of the explanatory variables $x_n = (x_{n1}, \ldots, x_{nM})^T$ and the dependent variable $t_n (n \in (1, ..., N))$. The likelihood is

$$p\left(t|x, w, \beta\right) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | w^T x_n, \beta^{-1}\right),$$

where w is the regression coefficient and β is the noise precision parameter. In order to simplify the discussion, we assume that the noise precision parameter β is known and consider a conjugate Gaussian prior distribution over w, $p\left(w|\alpha\right) = \mathcal{N}\left(w|0, \alpha^{-1} I\right)$. α determines the extent of shrinkage. When α is known, the posterior distribution of w follows a normal distribution. To allow for the uncertainty with this extent, a gamma prior distribution is introduced over α, $p\left(\alpha\right) = Gam\left(\alpha|a_0, b_0\right)$. In this case, the posterior distribution cannot be expressed explicitly.

In the variational framework, the mean field representation of w and α is

$$q\left(w, \alpha\right) = q\left(w\right) q\left(\alpha\right)$$

A practical variational distribution is

$$\begin{aligned} q\left(w\right) &= \mathcal{N}\left(w|m_N, S_N\right) \\ q\left(\alpha\right) &= Gam\left(\alpha|a_N, b_N\right). \end{aligned}$$

The joint density and the mean-field family are combined in order to form the ELBO for Bayesian ridge regression model. It is a function of the variational parameters m_N, S_N and a_N, b_N.

$$\begin{aligned} &\mathcal{L}\left[q(w, \alpha|m_N, S_N, a_N, b_N)\right] = \mathrm{E}_q\left[\log\left\{p\left(w, \alpha, t\right)\right\}\right] \\ &-\mathrm{E}_q\left[\log\left\{q(w, \alpha|m_N, S_N, a_N, b_N)\right\}\right] \\ &= \mathrm{E}_q\left[\log p\left(t|w\right)\right] + \mathrm{E}_q\left[\log p\left(w|\alpha\right)\right] + \mathrm{E}_q\left[\log p\left(\alpha\right)\right] \\ &-\mathrm{E}_q\left[\log q\left(w|m_N, S_N\right)\right] - \mathrm{E}_q\left[\log q\left(\alpha|a_N, b_N\right)\right]. \end{aligned}$$

Using the coordinate ascent algorithm, we update each variational parameter in turn as follows:

- update m_N, S_N: Given the values $a_N = a_N^0, b_N = b_N^0, m_N = m_N^0, S_N = S_N^0$, the value of m_N, S_N is updated as

$$\begin{aligned} m_N &= \beta S_N^0 X^T t \\ S_N &= \left(\mathrm{E}_q\left[\alpha\right] + \beta X^T X\right)^{-1} I = \left(\frac{a_N^0}{b_N^0} + \beta X^T X\right)^{-1} I, \end{aligned}$$

  where X is the design matrix $(x_1, \ldots, x_N)^T$.

- update a_N, b_N: Given the values $a_N = a_N^0, b_N = b_N^0, m_N = m_N^0, S_N = S_N^0$, the value of a_N, b_N is updated as

$$\begin{aligned} a_N &= a_0 + \frac{M}{2} \\ b_N &= b_0 + \frac{1}{2}\mathrm{E}_q\left[w^T w\right] = b_0 + \frac{1}{2}\left({m_N}^{0^T} {m_N}^0 + {S_N}^0\right) \end{aligned}$$

Using the optimized value of the parameters, the posterior means are estimated as:

$$\begin{aligned} \mathrm{E}_q\left[w\right] &= m_N \\ \mathrm{E}_q\left[\alpha\right] &= \frac{a_N}{b_N} \end{aligned}$$

5.3.2 Variational Inference of Topic Model

The second example is the simplest topic model - Latent Dirichlet Allocation (LDA). In case of LDA model, the observed variables are the words, organized into documents. The input data w_{dn} denote the n^{th} word in the d^{th} document. Across a collection, the documents share the same mixture components which are called topics β_k for $k \in (1, ..., K)$. A vector of topic proportions θ_d for $d \in (1, ..., D)$ describes the degree to which each document exhibits those topics. The LDA model assumes Dirichlet priors for both β_k and θ_d:

$$p\left(\beta_k|\eta\right) = Dir\left(\beta_k|\eta, \ldots, \eta\right)$$
$$p\left(\theta_d|\alpha\right) = Dir\left(\theta_d|\alpha, \ldots, \alpha\right).$$

The topic assignment z_{dn} denotes each word in each document which is assumed to have been drawn from a single topic. Therefore, the topics, topic proportions and topic assignments are latent variables. The posterior distribution is written generally as

$$p(\beta, \theta, z \mid w) = \frac{p\left(\beta, \theta, z, w\right)}{\int_\beta \int_\theta \sum_z p\left(\beta, \theta, z, w\right).}$$

However, the denominator is computationally infeasible.

In the variational framework for topic models, we consider firstly the variational distributions for latent variables. The mean-field variational family contains approximate posterior densities of the form

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q\left(\beta_k|\lambda_k\right) \prod_{d=1}^{D} q\left(\theta_d|\gamma_d\right) \prod_{n=1}^{N} q\left(z_{d,n}|\phi_{d,n}\right).$$

The factors $q\left(\beta_k|\lambda_k\right)$ and $q\left(\theta_d|\gamma_d\right)$ are the Dirichlet distributions on the k^{th} topic with global per-topic Dirichlet parameter λ_k and the d^{th} document with local per-document Dirichlet parameter γ_d. The factor $q\left(z_{d,n}|\phi_{d,n}\right)$ is a multinomial distribution on the n^{th} observation's topic assignment; its local assignment probabilities are a K-vector $\phi_{d,n}$. We construct the general ELBO for LDA model by combining the joint density and the mean-field variational family,

$$\begin{aligned}
&\mathcal{L}\left[q(\beta, \theta, z|\lambda, \gamma, \phi)\right] = \mathrm{E}_q\left[\log\left\{p\left(\beta, \theta, z, w\right)\right\}\right] \\
&-\mathrm{E}_q\left[\log\left\{q(\beta, \theta, z|\lambda, \gamma, \phi)\right\}\right] \\
&= \mathrm{E}_q\left[\log p\left(w|\beta, \theta, z\right)\right] + \mathrm{E}_q\left[\log p\left(z|\beta, \theta\right)\right] \\
&+\mathrm{E}_q\left[\log p\left(\theta|\alpha\right)\right] + \mathrm{E}_q\left[\log p\left(\beta|\eta\right)\right] \\
&-\mathrm{E}_q\left[\log q\left(z|\phi\right)\right] - \mathrm{E}_q\left[\log q\left(\theta|\gamma\right)\right] - \mathrm{E}_q\left[\log q\left(\beta|\lambda\right)\right].
\end{aligned}$$

With the complete conditionals, we now use the coordinate ascent variational inference algorithm which iterates between updating each local variational parameter and updating the global variational parameter:

- update the local variational parameter: Given the values of $\phi_{dn}^k = \phi_{dn}^{k0}, \gamma_{dk} = \gamma_{dk}^0, \lambda_{kn} = \lambda_{kn}^0$, the values of ϕ_{dn}^k and γ_{dk} are updated as

$$\begin{aligned}
\phi_{dn}^k &\propto \exp\left(E_q\left[\log\left(\theta_{dk}\right)\right] + E_q\left[\log\left(\beta_{kn}\right)\right]\right) \\
&= \exp\left(\Psi\left(\gamma_{dk}^0\right) - \Psi\left(\sum_{k'=1}^{K} \gamma_{dk'}^0\right)\right. \\
&\quad \left. +\Psi\left(\lambda_{kn}^0\right) - \Psi\left(\sum_{n'=1}^{N} \lambda_{kn'}^0\right)\right) \\
\sum_{k=1}^{K} \phi_{dn}^k &= 1 \\
\gamma_{dk} &= \alpha + \sum_{n=1}^{N} E_q\left[z_{dn}^k\right] = \alpha + \sum_{n=1}^{N} \phi_{d,n}^{k0}.
\end{aligned}$$

Here, $\Psi(.)$ is the digamma function, the first derivative of the log Gamma function.

- update the global variational parameter: Given the values of $\phi_{dn}^{k} = \phi_{dn}^{k0}, \gamma_{dk} = \gamma_{dk}^{0}, \lambda_{kn} = \lambda_{kn}^{0}$, the values of λ_{kn} are updated as

$$\lambda_k = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} E_q\left[z_{dn}^{k}\right] w_{dn} = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} \phi_{dn}^{k0} w_{dn}.$$

5.4 Variational Inference of CAT Model

Using results from the simplest family of distributions as mean-field variational approximations (Blei *et al.* 2006; Hoffman *et al.* 2013), each variable in the CAT-Poisson model is independent and governed by its own variational parametric distribution. Moreover, we consider truncated stick-breaking representations, proposed previously by Blei *et al.* (2006), for only the variational distributions. The truncated level or the largest number of categories K_{max} can be freely chosen. The family of variational distributions in the CAT-Poisson model can be written as follows:

$$\begin{aligned}
& q(\Xi, z, V, \pi, l, r|\Theta) \\
= & \prod_{j} q\left(l_j|\gamma_j, \gamma_j^{'}\right) \times \prod_{i} q\left(r_i|\zeta_i, \zeta_i^{'}\right) \\
& \times \prod_{k=1}^{K_{\max}} \prod_{a=1}^{20} q\left(\pi_a^k|\lambda_a^k\right) \times \prod_{k=1}^{K_{\max}} q\left(V_k|\vartheta_k, \vartheta_k^{'}\right) \\
& \times \prod_{i} \prod_{k=1}^{K_{\max}} q\left(z_i^k|\phi_i^k\right) \times \prod_{ij} q\left(n_{ij}|\omega_{ij}\right) \\
& \times \prod_{k=1}^{K_{max}} \prod_{a=1}^{20} q\left(w_a^k|\iota_a^k\right)
\end{aligned} \tag{3}$$

where

$$\begin{aligned}
q\left(l_j|\gamma_j, \gamma_j^{'}\right) &= Gamma\left(l_j|\gamma_j, \gamma_j^{'}\right) \\
q\left(r_i|\zeta_i, \zeta_i^{'}\right) &= Gamma\left(r_i|\zeta_i, \zeta_i^{'}\right) \\
q\left(\pi_a^k|\lambda_a^k\right) &= Dirichlet\left(\pi_a^k|\lambda_a^k\right) \\
q\left(V_k|\vartheta_k, \vartheta_k^{'}\right) &= Beta\left(V_k|\vartheta_k, \vartheta_k^{'}\right) \\
q\left(z_i^k|\phi_i^k\right) &= Multinomial\left(z_i^k|\phi_i^k\right) \\
q\left(n_{ij}|\omega_{ij}\right) &= Poisson\left(n_{ij}|\omega_{ij}\right) \\
q\left(w_a^k|\iota_a^k\right) &= Multinominal\left(w_a^k|\iota_a^k\right).
\end{aligned} \tag{4}$$

$\Theta = \left\{\gamma_j, \gamma_j^{'}, \zeta_i, \zeta_i^{'}, \lambda_a^k, \vartheta_k, \vartheta_k^{'}, \phi_i^k, \omega_{ij}, \iota_a^k\right\}$ are free variational parameters. To guarantee the tractability of computing the expectations of variational distributions, we choose variational distributions from exponential families (Wainwright *et al.* 2008).

To estimate each variational parameter in the CAT-Poisson model (3,4), we consider dividing the set of variational variables into two subgroups - global variables $[\Phi_g = (\Xi, \pi, l, r)]$ and local variables $[\Phi_l = (V, z)]$. The local variational variables (V, z) are per-data-point latent variables. The kth local variable V_k is unit length of kth stick in stick-breaking representation which is used to make the infinite vector of mixing proportions. The ith local variable z_i^k of the mixture component represents the allocation situation of site i of alignment of amino-acid sequences. Each local variable $\left(V_k, z_i^k\right)$ are governed by "local variational parameters" $\left[\Theta_l = \left(\vartheta_k, \vartheta_k^{'}; \phi_i^k\right)\right]$. Bishop (2006) has proposed coordinate ascent algorithm for solving the optimization problem of these variables. The coordinate ascent algorithm tries to find the local optimum of the ELBO

by optimizing each factor of the mean field variational distribution, while fixing the others. The optimal $q(z)$ and $q(V)$ are then proportional to the exponentiated expected log of the the joint distribution,

$$q^*(z) \propto \exp\left(E_{|z}\left[\log p(\Xi, V, z, \pi, l, r)\right]\right) + const$$
$$q^*(V) \propto \exp\left(E_{|V}\left[\log p(\Xi, V, z, \pi, l, r)\right]\right) + const.$$

Here, $E_{|z}$ and $E_{|V}$ denote expectations with respect to the variational distributions of all the variables except for z or V. The global variables Φ_g potentially control any of the data. These variables are governed by the "global variational parameters" $\left[\Theta_g = \left(\gamma, \gamma', \zeta, \zeta', \lambda, \omega, \iota\right)\right]$. The coordinate ascent algorithm iterates t times to update local variational parameters based on mapping data,

$$\Theta_l = \mathrm{E}_{\Theta_g}\left[\eta(\Phi, \Xi)\right]$$

where $\eta(.)$ are the natural parameters.

To estimate each global variational parameter in the CAT-Poisson model, we use the stochastic variational inference (SVI) algorithm to optimize the lower bound in Equation (2) (Hoffman *et al.* 2013). The stochastic variational algorithm is based on stochastic gradient ascent, the noisy realization of the gradient. The natural gradients (**?**) are adopted to account for the geometric structure of probability parameters (Robbins and Monro 1951). Importantly, natural gradients are easy to compute and give faster convergence than standard gradients. The SVI repeatedly subsamples the data, updates the values of the local parameters based on the subsampled data, and adjusts the global parameters in an appropriate way. Such estimates can guarantee algorithms to avoid shallow local optima of complex objective functions.

In our setting, we sample a mapping data point Ξ_n at each iteration, and compute the conditional natural parameters for the global variational parameters given N replicates of Ξ_n. Then, the noisy natural gradients are obtained. By using these gradient, we update Θ_g at each t iteration (with step size ρ_t)

$$\widehat{\nabla_{\Theta_g}\mathcal{L}} = prior + N\left\{E_{\Theta_l}\left[t(\Phi_n, \Xi_n), 1\right]\right\} - \Theta_g$$
$$\Theta_g^{(t)} = \Theta_g^{(t-1)} + \rho_t \widehat{\nabla_{\Theta_g}\mathcal{L}}$$

where $t(.)$ denote the sufficient statistics.

Based on the subsampling techniques, this procedure reduces the computational burden by avoiding the expensive sums in the above lower bound. The SVI algorithm thus significantly accelerates the variational objective analysis of the large database. Applying the previously proposed SVI framework (Hoffman *et al.* 2013), we can separate the computational cycle into the following steps:

1. Sample amino acid data from the whole set of input data.
2. Estimate how each site is assigned to a category, on the basis of observational data and the current approximation of variational parameters.
3. Update variational parameters
    - Local parameters are assignment variables, and breaking proportions.
    - Global parameters are equilibrium frequency profile, branch length, and rate across sites.

The lower bound of the data in terms of the variational parameters is specifically described in the Supplementary Material. Mathematical details of the variational objective function and computational methods of noisy derivatives and updating of variational parameters are also explained in that section.

5.5 Parallelization and Tree Topology

To parallelize the algorithm at the single machine level and thus reduce runtimes, we adopted the MPI parallelization of the PhyloBayes MPI program (Lartillot *et al.* 2013). Specifically, we use one master process for dispatching computational tasks and collecting and summing results, and with multiple slave processes executing the orders and returning all essential information to the master. This parallel strategy helps to equally divide the computational burden among slaves.

In addition, a partial Gibbs sampling algorithm for subtree pruning and regrafting (SPR) is adopted to update the tree topology (Lartillot *et al.* 2013). In a parallel environment, the task of the master process is to randomly select a subtree for pruning and send this information to all slaves. The task of each slave process is to update the conditional likelihood vectors of each resulting topology and the complete scan of all possible regrafting points. One single log likelihood for each regrafting point is arranged into an array and sent back to the master process. All arrays are collected and summed and lastly the Gibbs-sampling decision rule is finally applied to select regrafting position.

5.6 Data Sets

Three real data sets were used for our computational experiments. Data set A was a mitochondrial data set which consisting of 33 proteins, 6,622 amino acid positions from 13 species. Data set B was a plastid data set which composed of 50 plastidencoded proteins, 10,137 amino acid positions from 28 species. A total of 13% and 5% amino acid positions were missing from the mitochondrial and plastid data sets, respectively (Rodríguez-Ezpeleta *et al.* 2006; Lartillot *et al.* 2013). Finally, data set C was a more challenging and larger complete set of mitochondrial protein sequences derived from, a large alignment of EST and genome data, which consists of 197 genes, a total of 38,330 amino-acid positions from 66 species and with 30% missing data, is constructed by (Philippe *et al.* 2011).

C++ code for the variational inference version of the CAT model to perform computational experiments with these data sets is available at https://github.com/tungtokyo1108/.

6 Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online

7 Acknowledgments

This study was supported by Grant-in-Aid for Scientific Research (B) 16H02788 from the Japan Society for the Promotion of Science. We thank Edanz Group (www.edanzediting.com/ac) for editing the English text of a draft of this manuscript.

References

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Blei, D. M., Jordan, M. I., *et al.* 2006. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1): 121–143.

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

Ferguson, T. S. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Goldman, N., Thorne, J. L., and Jones, D. T. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of molecular biology*, 263(2): 196–208.

Gopalan, P., Hao, W., Blei, D. M., and Storey, J. D. 2016. Scaling probabilistic models of genetic variation to millions of humans. *Nature genetics*, 48(12): 1587.

Gopalan, P. K. and Blei, D. M. 2013. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36): 14534–14539.

Green, P. J. and Richardson, S. 2001. Modelling heterogeneity with and without the dirichlet process. *Scandinavian journal of statistics*, 28(2): 355–375.

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7): 910–917.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1): 1303–1347.

Ishwaran, H. and James, L. F. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453): 161–173.

Jones, D. T., Orengo, C. A., and Thornton, J. M. 1996. Protein folds and their recognition from sequence. *Protein structure prediction a practical approach*, pages 173–204.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233.

Koshi, J. and Goldstein, R. 1998. Models of natural mutations including site heterogeneity. *Proteins*, 32(3): 289–295.

Lartillot, N. 2006. Conjugate gibbs sampling for bayesian phylogenetic models. *Journal of computational biology*, 13(10): 1701–1722.

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.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. 2013. Phylobayes mpi: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62(4): 611–615.

Nielsen, R. 2002. Mapping mutations on phylogenies. *Systematic biology*, 51(5): 729–739.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428): 96.

Papaspiliopoulos, O. and Roberts, G. O. 2008. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1): 169–186.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., and Telford, M. J. 2011. Acoelomorph flatworms are deuterostomes related to xenoturbella. *Nature*, 470(7333): 255.

Raj, A., Stephens, M., and Pritchard, J. K. 2014. faststructure: variational inference of population structure in large snp data sets. *Genetics*, 197(2): 573–589.

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rodríguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B., and Melkonian, M. 2006. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of mesostigma in the streptophyta. *Molecular Biology and Evolution*, 24(3): 723–731.

Thorne, J. L., Goldman, N., and Jones, D. T. 1996. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13(5): 666–673.

Wainwright, M. J., Jordan, M. I., *et al.* 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305.