Improving Phylogenetic Analyses by Incorporating Additional Information from Genetic Sequence Databases

Li-Jung Liang 1, Robert E. Weiss 2, Benjamin Redelings 3, and Marc A. Suchard 2, 4

1 Department of Medicine Statistics Core, UCLA School of Medicine, 2 Department of Biostatistics, UCLA School of Public Health, and 4 Departments of Biomathematics and Human Genetics, UCLA School of Medicine, Los Angeles, CA 90095, USA, and 3 Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606, USA

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: Statistical analyses of phylogenetic data culminate in uncertain estimates of underlying model parameters. Lack of additional data hinders the ability to reduce this uncertainty, as the original phylogenetic data set is often complete, containing the entire gene or genome information available for the given set of taxa.

Informative priors in a Bayesian analysis can reduce posterior uncertainty; however, publicly available phylogenetic software specifies vague priors for model parameters by default. We build objective and informative priors using hierarchical random effects models that combine additional data sets whose parameters are not of direct interest but are similar to the analysis of interest.

Results: We propose principled statistical methods that permit more precise parameter estimates in phylogenetic analyses by creating informative priors for parameters of interest. Using additional sequence data sets from our lab or public databases, we construct a fully Bayesian semiparametric hierarchical model to combine data sets. A dynamic iteratively reweighted Markov chain Monte Carlo algorithm conveniently recycles posterior samples from the individual analyses. We demonstrate the value of our approach by examining the insertion-deletion (indel) process in the enolase gene across the Tree of Life using the phylogenetic software BALI-PHY; we incorporate prior information about indels from 82 curated alignments downloaded from the BALiBASE database.

Contact: liangl@ucla.edu

1 INTRODUCTION

In an experimental study, it is often possible to collect further data to more accurately estimate model parameters. In contrast, in comparative analyses of molecular sequence data, sequences are finite and typically complete, and more data cannot be collected. Additional sequences cannot be directly included in the analysis because that changes the inference target. Decreasing estimation standard errors in phylogenetic analysis requires a radically different approach with which to inject additional information into the analysis.

Bayesian methods can improve inference accuracy by incorporating proper informative prior distributions into the analysis (Gelman et al., 2003; Carlin and Louis, 2008). However, concerns often exist about the appropriateness of any given piece of prior information (Box and Tiao, 1992). The ability to generate appropriate prior distributions for phylogenetic parameters would be a useful mechanism to reduce estimation error (Alfaro and Holder, 2006).

Many phylogenetic conclusions are indeed sensitive to prior information (Rannala, 2002; Zwickl and Holder, 2004). For example, Yang and Rannala (2005) find that the posterior distribution of the phylogenetic tree is sensitive to the prior specification for internal branch lengths. When there is sensitivity, it is preferable to have priors that independent observers will recognize as encoding appropriate and relevant information. Internal branch length priors based on scientific information will minimize concern regarding this influence (Kolaczkowski and Thornton, 2007).

Subjective priors form the basis of traditional Bayesian proper prior construction. One elicits these priors from the statistician or other expert researchers (for example Bedrick et al., 1996, 1997) or studies reported in the literature. In the case of phylogenetic analyses, prior information is often available in the form of previous analyses of similar data sets, either in the literature or from a biologist’s own laboratory. For example, further sequence data from similar taxa or genes is often available.

An objective approach to proper prior construction begins by building a Bayesian hierarchical model involving the data set(s) of interest and additional similar data sets whose analyses may not be of immediate interest. Biologists can find or construct data sets similar to their own data either from their lab or from proliferating publicly available sequence databases. Candidate databases include the benchmark alignments database (BALiBASE, Thompson et al., 1999; http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/, version 1) described in Section 2, and HIV sequence databases at the Los Alamos National Laboratory for researchers who study HIV sequences and drug resistance (http://www.hiv.lanl.gov/content/index).

For aligned or alignable sequences, biologists often use publicly available software to estimate parameters describing the evolutionary process and the unknown phylogenies that give rise to the sequences. These programs fit complex Bayesian models; running time can be substantial. Priors are relatively vague, at least as the default option, as in MrBAYES (Huelsenbeck and Ronquist, 2001) and BALI-PHY (Suchard and Redelings, 2006). These programs are designed to fit single data sets and are not designed to fit a hierarchical model containing multiple data sets. Informative priors may or may not be easily set in the software, depending on implementation specifics.

In this paper we introduce easy-to-use methods that allow biologists to construct informative priors for particular analyses using additional information either from their own lab or public sequence databases. We assume that K (one or a small number) of data sets are to be analyzed using the publicly available software.

The first step involves the construction of an additional N related data sets of sequences. These additional data sets are similar in the sense that the K + N data sets may be thought of as random samples from the same sample space of data sets, (exchangeable in Bayesian parlance) possibly after adjusting for covariates. Secondly, we analyze each of the N + K data sets independently using the
chosen software and obtain Markov chain Monte Carlo (MCMC) samples from the marginal posterior distributions of the phylogenetic model parameters of interest. We refer to each of these $N + K$ analyses as an individual analysis. In the third step, we propose to fit a semiparametric Bayesian hierarchical random effects model that combines the $N + K$ individual analyses.

Our hierarchical model exploits a mixture of Dirichlet processes (DPs) (Sethuraman, 1994; Escobar and West, 1995) for the prior of the random effects, allowing for equality in parameters across some, but not necessarily all, analyses. In effect, the hierarchical model with mixture DP prior simultaneously enables us to borrow strength from the additional analyses to improve individual inferences and protects against pooling the analysis of interest with analyses lacking information about the parameters of interest. The hierarchical model then updates the individual analysis posterior given a single data set to a posterior given the full set of $N + K$ data sets. Our inference tools for this step avoid rewriting the public domain software yet jointly fit all data sets simultaneously by reusing the available MCMC output generated from the individual analyses.

This third step can deliver a convenient proper prior that may be reused to update the existing software for future analyses. We approximate the posterior predictive distribution (PPD) of the parameters of interest; this distribution is exactly the correct prior for a new analysis. When feasible, we recommend incorporating the posterior predictive distributions as an option into the public software for use in analyzing future data related sets of interest. Alternatively, we show how to update a future analysis without modifying the software or rerunning our software.

Our hierarchical prior model has several distinct features; first, parameters of interest can be univariate, bivariate or higher dimensional. Second, our approach is practicable whether or not we can make software modifications. Third, additional data sets may be directly exchangeable with the data sets of interest, or may be exchangeable after adjustment for covariates. Fourth, rather than apply the nonparametric DP prior directly to the random effects, we instead apply it to the hyperparameters of the random effects distribution for additional flexibility. Fifth, we exploit a dynamic iteratively reweighting MCMC algorithm (DyIRMA) proposed by Liang and Weiss (2007) that replaces the slow individual analysis updates with efficient re-sampling procedures.

Our approach is in the spirit of Empirical Bayes (EB) (Robbins, 1956; Morris, 1983; Efron, 1996; Carlin and Louis, 2000) in the sense that the informative (empirical) priors derive from data rather than from a priori assumptions. A key difference is that we start with $K$ data sets of interest, then gather $N$ similar data sets whose output may not be of direct interest. Thus, our approach is fully Bayesian and makes use of hierarchical models, but the primary goal is to estimate $K$ random effects, rather than for the more common purpose of estimating parameters of the hyper prior. It may be that given the additional $N$ data sets, the investigators may switch attention to the hyperparameter estimates, but that is not our intent; we remain focused on the $K$ original analyses.

2 DATA AND APPLICATION

We wish to examine the insertion-deletion (indel) process in the enolase gene. Our motivating data set consists of 37 aligned enolase sequences identified in Baptiste and Philippe (2002).

Indels are not directly observed in sequences; rather, one infers their existence through sequence alignment. Appropriate inference about where indel events occur depends on correctly specifying indel process parameters. Sequence alignment in turn depends on the tree relating the sequences; yet inference about the phylogenetic tree depends on the alignment. Considerable uncertainty exists regarding the evolutionary tree relating taxa across the Tree of Life in turn causing difficulty in alignment. To control for this uncertainty and protect against bias (Lake, 1991), we use a statistical model that jointly infers alignment and phylogeny (Redelings and Suchard, 2005). This joint framework uses a hidden Markov model (HMM) to characterize the indel process, parameterized by an indel opening rate $\lambda$ and expected additional indel length $l = \epsilon/(1 - \epsilon)$ (Redelings and Suchard, 2007), which is one less than the mean indel length.

Our specific scientific objective is to learn about the bivariate distribution of $(\log \lambda, \log l)$ in the enolase genes of a specific set of sequences. Unfortunately, the enolase data are sparse, causing uncertain inference for the key parameters $(\log \lambda, \log l)$. We apply our approach to increase the precision of inference.

First, we identify multiple additional data sets that when properly analyzed provide useful information about the parameters of interest. We turn to the BALiBASE database to furnish these additional data sets. BALiBASE is a well developed, publicly available database of curated alignments for evaluation of sequence alignment programs. The BALiBASE first reference set contains 82 reliable alignments, of 3 to 5 equi-distance sequences each, where the percent identity between two sequences lies within an identified range. The diversity level can be low (<25%), medium (25%−35%), and high (>35%). The alignment length can be short, medium or long.

We analyze individual data sets with the software package BALi-PHY. BALi-PHY uses an MCMC algorithm to sample from the joint posterior distribution of alignment, alignment parameters, phylogenetic tree and phylogenetic parameters given a set of sequences (Suchard and Redelings, 2006). Under the joint alignment and phylogeny model, the computational complexity of fitting a single data set is large; simultaneously fitting multiple data sets is currently impractical. Our DyIMRA algorithm side-steps this severe limitation.

To better illustrate the value of our approach and algorithm, we partitioned the 37 enolase gene sequences into 7 separate data sets containing 4-7 taxa each. These 7 data sets in turn provide 7 similar illustrations of our approach rather than a single example. Further, this demonstrates the particular benefit of our approach for small data sets.

3 METHODOLOGY

For simplicity, we assume $K$ data sets $Y_1, \ldots, Y_K$ of specific interest with corresponding parameters $\mu_1, \ldots, \mu_K$ of interest. We collect an additional $N$ data sets $Y_{K+1}, \ldots, Y_{K+N}$ that we consider similar to the data sets of interest. The $N$ data sets are not of interest at the current time and are collected only to increase the precision of parameters $\mu_1, \ldots, \mu_K$.

3.1 Individual Analyses

Define $|[X]$ and $[X|Z]$ to be the densities of random variable $X$ and the conditional density of $X$ given $Z$ respectively. Each data set $Y_i$, $i = 1, \ldots, K + N$ is assumed drawn independently from probability density $[Y_i | \mu_i, \theta, z_i]$, where parameter of interest $\mu_i$ is a $D$-dimensional vector, $\theta_i$ is a vector of nuisance parameters: additional parameters not of interest, and $z_i$ are covariates that describe known differences between data sets. Parameters of interest $\mu_i$ have different values across data sets. In
our situation, \(D = 2\) with \(\mu_i = (\log \lambda_i, \log \kappa_i)\). The nuisance parameters \(\theta_i\) may not be comparable across data sets. For example, the alignment of a set of molecular sequences or the tree relating those sequences is meaningful for those sequences and is not relevant for other sequences.

We use the public domain Bayesian software BAII-PHY to separately analyze each \(Y_i\) with sampling model \([Y_i|\mu_i, \theta_i, x_i]\) and prior \([\mu_i|\phi = \phi_0][\theta|\mu_i]\), where \(\phi\) is a vector of prior parameters, which, when we fix \(\phi = \phi_0\), we recover the prior in the software. To extract maximal information, we fix the alignments, but not the indel process, in the additional \(N\) BAliBASE data sets to their curated estimates. For the \(K\) data sets of interest, we have no prior information and estimate their alignments as part of their analyses. When analyzed independently, the \(K + N\) data sets give rise to \(K + N\) separate posteriors \([\mu_i, \theta_i|Y_i, x_i, \phi = \phi_0]\) which we call the individual posteriors.

### 3.2 Semiparametric Random Effects Model

We now combine information from the individual posteriors using our hierarchical model to improve inference on \(\mu_i\). To do this, we specify a joint posterior for the entire vector \(\mu = (\mu_1, \ldots, \mu_{K+N})\)' given the complete data \(Y = (Y_1, \ldots, Y_{K+N})\).

To pool information across data sets, we model individual \(\mu_i = (\mu_{i1}, \ldots, \mu_{id})\)' as coming from a \(D\)-variate normal (MVN)\(D\)

\[
\mu_i|\alpha, \beta, \Sigma \sim \text{MVN}_D \left( \begin{pmatrix} x_{i1}\alpha_1 + \beta_1 \\ \vdots \\ x_{id}\alpha_d + \beta_d \\ \end{pmatrix}, \Sigma \right),
\]

(1)

given a \(Q + 1\) vector of covariates \(x_{id} = (1, x_{i1}, \ldots, x_{iQ})\)' including the intercept, unknown regression coefficients \(\alpha_d = (\alpha_{0d}, \ldots, \alpha_{Qd})\)', a \(D \times D\) diagonal covariance matrix \(\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)\), and unknown data-set-specific random effects \(\beta_d\) for each dimension \(d = 1, \ldots, D\) of \(\mu_i\). In our example, we have \(Q = 4\) covariates, two indicators each for sequence length (median and long) and identity (median and high). We use these covariates to predict both \(\mu_{i1}\) and \(\mu_{i2}\) for a total of \(Q + 1 = 5\) regression coefficients \(\alpha\) per parameter of interest. In general, each parameter can have different predictors, this makes for an increase in notational complexity of one additional subscript and an increase in data management issues but it presents no theoretical difficulties. The multivariate normal assumption is assumed to hold possibly after transformation, as in our case where we take logs of both parameters before modeling.

To complete the hierarchical specification of (1), we merge all fixed-effects coefficients into a \((Q + 1) \times D\) matrix \(\alpha\) with columns \(\alpha_1, \ldots, \alpha_D\). Each row \(\alpha_{(q)}\) of \(\alpha\) corresponds to the same covariate \(q = 1, \ldots, Q + 1\) across the dimensions of \(\mu_i\). We also form \(\beta_i = (\beta_{i1}, \ldots, \beta_{iD})\). To relax parametric assumptions on \(\beta_i\), we envision an unknown, discrete probability distribution \(P\) as the prior for the \(\beta_i\) and model \(P\) with a Dirichlet process, such that

\[
\begin{align*}
\alpha_{(q)}|\eta_q, \nu_q \sim \text{MVN}_D(\eta_q, \nu_q), \\
\beta_i|P \sim P, P \sim \text{DP}(M, P_0), \\
\Sigma = \text{MVN}_D(0, \Sigma_0), \\
\sigma_d^{-2} \sim \text{Gamma}(c_d, s_d), \\
\Sigma_0^{-1} \sim \text{Wishart}_D(a_0, R_0),
\end{align*}
\]

where \(P_0\) is the mean of the DP generating \(P\) with parameter \(M\). We fix \(M, \alpha_0, R_0, c_d, s_d, \eta_q = (\eta_{q1}, \ldots, \eta_{qQ})\), and \(V = (V_1, V_2)\) as a priori known constants of the hyper-prior. We estimate the unknown hyper-parameters \(\phi = (\alpha, \beta, \Sigma_0, \Sigma_1)\) as part of the model. A fixed value \(\phi_0\) of \(\phi\) is assumed to yield the priors used in the individual analyses, so that \([\mu_i = [\mu_i|\phi = \phi_0]\).

### 3.3 Computation

The public domain software is used to fit the individual models \([Y_i|\mu_i, \theta_i, x_i]\) with prior \([\mu_i|\phi = \phi_0][\theta|\mu_i]\). This software in turn produces posterior samples from \([\mu_i, \theta_i|Y_i]\). This model and resulting posterior is a complex function of \(\mu_i\) and the posterior does not have a closed form algebraic representation. We run each individual model \(i = 1, \ldots, K + N\), producing a Markov chain Monte Carlo (MCMC) sample from each posterior.

We do not wish to make the huge investment that would be necessary to reprogram this model and combine it with our hierarchical model (1) and (2). Furthermore, fitting the individual prior can take hours if not days to produce a single sample; re-running the \(K + N\) individual models jointly along with our hierarchical model could produce a problem that might well push past the limits of computational tractability. In the not uncommon situation where all \(K + N\) analyses might be of interest, for example when all sets of sequences come from the same lab, we would not want to rerun all the individual analyses, wasting the computer time already spent on each analysis. Our approach uses a divide and conquer approach to break the problem into the \(K + N\) individual models plus one hierarchical model that uses the results from the individual models resulting in more manageable programming and computational efforts.

To combine the results of the \(K + N\) individual posteriors, a conventional Gibbs (Gelfand and Smith, 1990) or Metropolis-Hastings (Tierney, 1994) sampling algorithm is not possible. Instead, Liang and Weiss (2007) propose DyIRMA to replace the difficult sampling of \(\mu\) with a weighted sampling of the MCMC realizations of \(\mu\) from the individual posteriors, eliminating the need for redundant reprogramming, and the computational cost of rerunning the original models.

Let \([\mu_i|\phi_0]\) and \([\theta|\mu_i]\) be the prior densities of \(\mu_i\) and \(\theta_i\) from the individual analyses. The densities \([\mu_i, \theta_i|Y_i, \phi_0]\) and \([\mu_i|Y_i, \phi_0]\) are the joint and marginal individual posterior densities for \((\mu_i, \theta_i)\) and \(\mu_i\), respectively. Let \([\mu_i|\phi]\) denote the prior density for \(\mu_i\) from our hierarchical model; we refer to this as the hierarchical prior for \(\mu_i\). In the hierarchical model, \([\mu_i|Y_i]\) and \([\mu_i|Y_i]\) are the hierarchical posteriors for \(\mu_i\) and \(\mu_i\); for \(i = 1, \ldots, K\) these are the quantities that we aim to estimate.

Integrating each joint posterior density \([\mu_i, \theta_i|Y_i, \phi_0]\) over its nuisance parameter vector \(\theta_i\) yields the individual marginal posterior \([\mu_i|Y_i, \phi_0]\). Dividing this density by the prior distribution \([\mu_i|\phi]\) used in the individual model and multiplying by the proposed hierarchical prior density \([\mu_i|\phi]\) is the individual contribution from each individual problem to our posterior. We multiply these individual contributions together and multiply them by the hierarchical prior density \(\phi\). The resulting hierarchical posterior is

\[
[mu_i|Y_i] \propto \int [phi] \prod_{1 \leq i \leq N} \left[ \frac{[\mu_i|\phi]}{[\mu_i|\phi_0]} \right] d\phi.
\]

(3)
The ratio

$$w^*_\alpha (\mu_i, \phi) = \frac{[\mu_i | \phi]}{[\mu_i | \phi_0]}$$  \hspace{1cm} (4)$$

is a weight function; hence, conditional on \(\phi\), the posterior of \(\mu_i\) under the hierarchical model is a reweighting by (4) of the original marginal individual posterior \([\mu_i | Y_i]\). Unconditional on \(\phi\), this statement is still true, \([\mu_i | Y]\) is a reweighting of \([\mu_i | Y_i]\).

The conventional Gibbs algorithm simulates the joint posterior distribution \([\mu, \phi | Y]\) by sampling \(\mu_1, \ldots, \mu_K, \phi\) sequentially from their full conditional distributions \([\mu_i | Y, \phi, \mu_{-i}] = [\mu_i | Y_i, \phi_i]\), where \(\mu_i\) is \(\mu\) without the \(i^{th}\) component, and \([\phi | Y, \mu]\). Instead, equation (3) suggests the DyIRMA framework. We update \(\mu_i\) by drawing a realization from the pre-generated MCMC samples of \([\mu_i | Y, \phi_0]\) weighted by \(w^*_\alpha (\mu_i, \phi)\). The reweighting occurs dynamically inside the Gibbs step and is a function of the current value of \(\phi\).

In detail, let \(\mu_i^{(m)}\) be the \(m^{th}\) MCMC sample value from \([\mu_i | Y, \phi_0]\) for \(m = 1, \ldots, M_i\), the number of MCMC samples resulting from the analysis of data set \(i\). Our algorithm replaces sampling from \([\mu_i | Y, \phi]\) with sampling from the weighted empirical density

$$[\mu_i | Y, \phi]_\epsilon = W_i^{-1} \sum_{m=1}^{M_i} w_{\phi_0}^{(m)} (\mu_i^{(m)}, \phi) \delta_{\mu_i^{(m)}} (\mu_i),$$  \hspace{1cm} (5)$$

where \(\delta_{\mu_i^{(m)}} (\cdot)\) denotes point mass at \(\mu_i^{(m)}\), and \(W_i = \sum_{m=1}^{M_i} w_{\phi_0}^{(m)} (\cdot, \phi)\). For each DyIRMA iteration \(t\), we sample \(\mu_i\) for each \(i\) from (5) given \(\phi = \phi^{(t-1)}\). We then generate \(\alpha^{(t)}\), \(\beta^{(t)}\), \(\Sigma^{(t)}\), and \(\sigma^{(t)}\) for \(d = 1, 2\) in sequential Gibbs steps from their full conditional distributions given \(\mu^{(t)}\). The samples drawn for \(\mu_i^{(t)}\) for \(i = 1, \ldots, K\) are the posterior samples for our \(K\) analyses of scientific interest under the full hierarchical model.

3.4 Posterior Predictive Density

We also use our method to generate an informative prior for the parameter of interest \(\mu_i\) for a future analysis with associated covariates \(x_i = (x_{i1}, \ldots, x_{id})'\). The prior we want to calculate is the PPD, \([\mu_i | Y]\), for a new data set \(Y_i\). Since \(\mu_i\), given both \(\phi\) and its mean \(g(\alpha) + \beta_i\), is independent of \(Y\), we have

$$[\mu_i | Y] = \int \int [\mu_i | g(\alpha), \beta_i, \phi] \times [\beta_i, \phi | Y] \, d\phi \, d\beta_i.$$  \hspace{1cm} (6)$$

Because the PPD (6) does not have a closed form, we use a Rao-Blackwell estimate (Gelfand and Smith, 1990) to estimate (6) as

$$\frac{1}{T} \sum_{t=1}^{T} [\mu_i | g(\alpha^{(t)}) + \beta_i^{(t)}, \phi^{(t)}],$$  \hspace{1cm} (7)$$

where \(\alpha^{(t)}\), and \(\phi^{(t)}\) are the \(t^{th}\) DyIRMA samples and \(T\) is the total number of samples after the burn-in period. The random effect \(\beta_i^{(t)}\) is new, we draw it from \([\beta_i | \phi] = \frac{1}{1+\lambda} \sum_{m=1}^{M} \delta_{\beta_i^{(m)}} + \frac{\lambda}{1+\lambda} P_0(\beta_i)\), where \(\delta_{\beta_i^{(m)}}\) is the distribution with point mass of the single point \(a\) (Blackwell and MacQueen, 1973; Bush and MacEachern, 1996). There is no additional programming effort to compute the posterior sample average (7). All we need is to add one step to the DyIRMA approach: generate \(\beta_i^{(t)}\) from \([\beta_i | \phi]\).

3.5 Incorporating Empirical Priors

When feasible to replace the vague prior for \(\mu_i\) used in the public phylogenetic software with the Rao-Blackwell estimate of the PPD (7), it is useful to approximate the density either parametrically or non-parametrically. The software is then updated to permit a parametric or nonparametric prior. To approximate a parametric form for the target density, we select a parametric family, such as the normal, \(t\) or Beta family, that is fairly close to the target density. Often visual inspection suffices to identify a satisfactory family. When the target density is clearly different from a known parametric form, we recommend approximating \([\mu_i | Y]\) nonparametrically. A kernel density estimator (KDE) \([\mu_i | Y]_{KDE}\) of \([\mu_i | Y]\) is

$$[\mu_i | Y] \approx \frac{1}{M} \sum_{m=1}^{M} K_s (\mu_i - \mu_i^{(m)}) \equiv [\mu_i | Y]_{KDE},$$  \hspace{1cm} (8)$$

where \(K_s (\cdot)\) is a density, usually continuous and symmetric and preferably continuously differentiable as well, with scale parameter \(s\).

The posterior density of \(\mu_i\) is \([\mu_i | Y_i, \phi] \propto [Y_i | \mu_i] [\mu_i | Y]_{KDE}\), with the approximated prior (8) used in the public phylogenetic software. In the situation where changing the public phylogenetic software is not feasible, we can of course still update \([\mu_i | Y_i]\) using our previously described procedure. If we have already run our software to produce \([\mu_i | Y]\) first, we next run the public domain software to produce MCMC samples \(\mu_i^{(m)}\) from \([\mu_i | Y_i], m = 1, \ldots, M_i\). Then \([\mu_i | Y_i, \phi] \propto [\mu_i | Y_i] [\mu_i | Y]_{KDE} (\mu_i | \phi)^{-1}\), where \(\mu_i | \phi\) is the prior used in the public domain software, we weight \(\mu_i^{(m)}\) with weights \(w^{(m)} = \frac{[\mu_i | Y_i] [\mu_i | Y]_{KDE} (\mu_i | \phi)^{-1}}{[\mu_i | Y_i] [\mu_i | Y]_{KDE} (\mu_i | \phi_0)^{-1}}\) giving a weighted posterior sample from \([\mu_i | Y_i, \phi]\).

3.6 Model Comparison

We discuss model comparisons in terms of the new data set \(Y_i\) with parameter \(\mu_i\). We wish to compare the posteriors \([\mu_i | Y_i]\) and \([\mu_i | Y_i]\) resulting from model \(M_0\) with the default prior and model \(M_0\) with the hierarchical prior, respectively. We propose two methods to evaluate which prior model provides better estimation. The first employs a Bayes factor (BF). The BF in favor of \(M_0\) against \(M_0\) for the single data set \(Y_i\) is

$$BF_{M_0} = \frac{[Y_i | M_0]}{[Y_i | M_0]} = \frac{\int [Y_i | \mu_i, \theta_i] [\mu_i | Y_i] [\theta_i | \mu_i] d\mu_i d\theta_i}{\int [Y_i | \mu_i, \theta_i] [\mu_i | \phi = \phi_0] [\theta_i | \mu_i] d\mu_i d\theta_i},$$

$$= \frac{\int [\mu_i | Y_i, \phi = \phi_0 | \mu_i]}{[\mu_i | \phi = \phi_0]} d\mu_i,$$  \hspace{1cm} (9)$$

when using the same priors for \(\theta_i\) in both models. Convenietly, we compute the Monte Carlo approximation of \(BF_{M_0}\) using the \(M_i\) posterior samples \(\mu^{(m)}\) from the analysis based on the model \([\mu_i | Y_i, \phi]\) with the default priors.

In addition to Bayes factors, we propose using the posterior expected loss (PEL) to compare the two models. Consider an expanded mixture model combining both \(M_0\) and \(M_0\) where the prior is a convex mixture \(v_0 \times [\mu_i | \phi = \phi_0] + (1 - v_0) \times [\mu_i | Y_i]\) of the public domain software prior \([\mu_i | \phi = \phi_0]\) and the hierarchical prior \([\mu_i | Y_i]\) for some choice of \(0 \leq v_0 \leq 1\). The mixture model assumes that one of the two priors is the correct prior; the parameter \(v_0\) can be interpreted as the prior probability of the public domain software prior having the correct prior. This is a family of models indexed by \(v_0\).
The resulting posterior density is a mixture \( v \times [\mu_i]_Y \circ \phi = \phi_0 \) + \( (1 - v) \times [\mu_i|Y, V] \), where \( 0 \leq v \leq 1 \) is the posterior probability of the model implemented in the public domain software. The posterior probability \( v \) is a monotone non-decreasing function of \( v_0 \), and can be calculated as part of the analysis. The posterior probability \( v = 0 \) when \( v_0 = 0 \) and \( v = 1 \) when \( v_0 = 1 \). The posterior mean \( E_v(\mu_i|Y) = v \times E_0(\mu_i|Y) + (1 - v) \times E_n(\mu_i|Y) \), and the PEL \( v \) of the posterior mean under squared error loss is, as a function of \( v \),

\[
v(1 - v) (E_0(\mu_i|Y) - E_n(\mu_i|Y))^2 + v \times V_0(\mu_i|Y) + (1 - v) \times V_n(\mu_i|Y),
\]

(10)

where \( V_0(\mu_i|Y) \) and \( V_n(\mu_i|Y) \) are posterior variances under the models \( M_0 \) and \( M_n \), respectively.

The PEL \( v \) equals \( V_0(\mu_i|Y) \) if \( v_0 = 0 \), and \( V_n(\mu_i|Y) \) if \( v_0 = 1 \). For \( v_0 \) in the interval \((0, 1)\), the PEL \( v \) is always greater than the smaller of \( V_0(\mu_i|Y) \) and \( V_n(\mu_i|Y) \). Therefore, using the posterior expected loss criterion, the best model out of the family of models indexed by \( v_0 \) overall occurs either at \( v_0 = v = 0 \) or at \( v_0 = v = 1 \) and we should prefer the model that gives us the smallest posterior variance. In our examples, the hierarchical model usually returns a smaller posterior variance and therefore a smaller posterior expected loss. If the \( N \) additional data sets are selected appropriately, the posterior resulting from the hierarchical model will be preferable to the individual posteriors for our \( K \) data sets of interest.

4 RESULTS

We illustrate our 3-step approach by first collecting prior information from the 82 BALiBASE data sets to help us to examine the indel process in play with enolase gene evolution. We use BALI-PHY to fit the individual analyses. The default priors on the log indel opening rate \( \mu_i \) and log indel extension \( \mu_e \) are independent. The default prior for \( \mu_i \) is a double exponential distribution, \( (2c)^{-1} \exp(-|\mu_i - b|/c) \), \( c > 0 \), with mean \( b = -5 \) and standard deviation \( \sqrt{2c} = 10 \). The default prior for \( \psi = \exp(\mu_e) \) is an exponential distribution, \( \exp(-d^{-1} \psi) \), with mean \( d = 10 \).

First, we pre-process the 82 data sets to meet the input requirements for BALI-PHY and then generate the individual bivariate posteriors \( [\mu_i|Y] \). For each analysis, we ran the MCMC sampler for 50,000 iterations and discard the first 10,000 samples as burn-in. We store every 40th sample yielding a total of 1000 realizations for each individual analysis. Then we run our \( K = 7 \) data sets of interest with the same sampling scheme. We use the samples from the independent analyses as input to our hierarchical model. Subsection 4.1 discusses the results from combining just the \( N \) prior analyses in a joint model. Subsection 4.2 presents the results of implementing a module for our empirical priors into BALI-PHY and using this prior to inform our examination of the enolase data sets.

4.1 Combining Analyses

We fit our hierarchical model with two data set-specific covariates, alignment length (short, medium, long) and percent-identity (low, medium, high), to combine the 82 individual prior analyses. We set the hierarchical hyper-prior constants as follows. Hyperprior mean \( \eta_i = (-5, 2.3) \) and covariance matrix \( V_0 = \text{diag}(4, 4) \), matching the prior means and variances for \( \mu_i \) and \( \mu_e \) used in BALI-PHY. Covariates \( \eta_q = (0, 0) \) and \( V_q = \text{diag}(1, 1) \), \( q = 2, \ldots, 5 \), are finite but relatively uninformative. The inverse covariance matrix \( \Sigma^{-1}_q \) is a diagonal matrix with elements \( \sigma_q^{-2}, d = 1, 2 \), that are gamma-distributed with mean 5, which is approximately the range of the individual posterior samples of \( \mu_i \), and variance 25, so that the prior mean is equal to the prior standard deviation. For \( \Sigma_0, \alpha_0 = 2 \) and \( R_0 = \text{diag}(5, 5) \) were chosen such that \( E(\Sigma_0) \) is approximately 30% larger than the variance of the individual posterior means of \( \mu_i \). Finally, we set the DP concentration parameter \( M = 5 \) using a formula suggested by Liu (1996) and generate 50,000 DyIRMA iterations with the first 5000 iterations discarded as burn-in.

Figure 1(a) shows a bivariate scatter plot of the 82 posterior means of \( \mu_i \) versus \( \mu_e \) from our hierarchical model. Covariate labels \( \circ, \triangle, \text{and} \times \) represent short, medium, and long length data sets, respectively. The horizontal line identifies the log expected indel length from our hierarchical model. Here \( \circ, \triangle, \text{and} \times \) represent short, medium, and long length data sets, respectively. Boxplots (graphical 5-number summary: minimum, 25%-percentile, median, 75%-percentile, and maximum) of variance reduction from the individual posteriors of \( \mu_i \) and \( \mu_e \). Each plot includes 9 entries; the clear, gray and solid boxes are for short, medium, and long length, respectively. The boxes labeled “L,” “M,” “H” are low, medium, and high levels of identity, respectively.

Fig. 1. (a) Scatter plot of the posterior mean of \( \mu_i \) (log indel opening rate) against the posterior mean of \( \mu_e \) (log expected indel length) from our hierarchical model. Here \( \circ, \triangle, \text{and} \times \) represent short, medium, and long length data sets, respectively. Boxplots (graphical 5-number summary: minimum, 25%-percentile, median, 75%-percentile, and maximum) of variance reduction from the individual posteriors of \( \mu_i \) and \( \mu_e \). Each plot includes 9 entries; the clear, gray and solid boxes are for short, medium, and long length, respectively. The boxes labeled “L,” “M,” “H” are low, medium, and high levels of identity, respectively.

...
percent reduction is 16% (range: 0.7% to 54%) and 17% (range: 0.7% to 54%) for

\[
\mu_{k1} = 1 - 8.64 \quad \text{and} \quad \mu_{k2} = 1 - 8.64.
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]

respectively. Two data sets (5 and 7) for

\[
\mu_{k1}, \quad \mu_{k2}, \quad \mu_{k3}, \quad \mu_{k4}, \quad \mu_{k5}, \quad \mu_{k6}, \quad \mu_{k7},
\]
using our approach furnish informative priors for the parameters of interest for future analyses of data sets having similar covariates.

We propose two performance evaluation methods for these priors, both of which are easy to compute. We combined each of seven data sets of interest with extensive prior information and then performed two phylogenetic analyses, one using the default priors for the indel process parameters in BALI-PHY and the other using the empirical prior generated using our approach. When data are sparse, our results show that careful prior choice becomes valuable and informative priors provide substantially more efficient estimates. Assuming that (1) the new data sets that researchers have are similar to the BALiBASE data sets, (2) researchers are interested in the indel model using BALI-PHY or other software, and (3) the prior model for the indel parameters is the same, then the prior generated here can be useful for analyzing their new data.

Our findings should encourage researchers to identify and gather as many sets of comparable sequence data as possible from publicly available sequence databases or research laboratories. Our approach can be applied to generate proper prior distributions of the phylogenetic parameters of interest. The priors can later be incorporated into existing software for use by others. When prior information is available, this helps researchers produce more accurate posterior estimates of the phylogenetic parameters of interest, as compared to using the default priors. In the case where empirical priors could not be incorporated directly into existing software, our approach can still be used to yield better posterior estimates by combining analyses of interest with additional prior analyses through dynamic re-weighting.

This re-weighting algorithm requires calculation of weights \( w_{	ext{Bayes}}(\mu_i, \phi) \) at each iteration for all individual posterior samples; this can become computationally expensive. Alternatively we can use a Metropolis-Hastings step (Tierney, 1994) sampling candidate \( \mu_{i}^{(t+)} \) at random from the unweighted samples. Both options perform optimally when the samples from the individual posteriors cover the entire posterior support of the \( \mu_i \). Liang and Weiss (2007) suggest another class of algorithms which approximate the individual posterior using a KDE. The KDE provides a smoothed version of the posterior using a KDE. The KDE provides a smoothed version of the posterior using a KDE. This can become computationally expensive. Alternatively we can use a Metropolis-Hastings step (Tierney, 1994) sampling candidate \( \mu_{i}^{(t+)} \) at random from the unweighted samples. Both options perform optimally when the samples from the individual posteriors cover the entire posterior support of the \( \mu_i \). Our approach also provides suggestions to software engineers. Software will be more useful and calculations can be made easier if future software engineers allow for KDE type priors as input into their programs and also allow for families of proper priors, including, but not limited to, normal, t and Beta priors as appropriate for individual parameters. Even if joint priors are restricted to products of independent priors, this will still allow for substantially improved inference.

ACKNOWLEDGEMENT

This work is supported by NIH grant GM086887. Liang is supported by a Biostatistics training grant for AIDS Research (ST32AI007307), Weiss is supported by the UCLA Center for AIDS Research, NIH grant AI28697 and Suchard is an Alfred P. Sloan Research Fellow and a John Simon Guggenheim Memorial Foundation Fellow.

REFERENCES


