Phylogenomic Analysis Using Bayesian Congruence Measuring

Dazhuo Li
Department of Computer Engineering and Computer Science
University of Louisville, Louisville, KY40217, USA

Eric C. Rouchka *
Department of Computer Engineering and Computer Science
University of Louisville, Louisville, KY40217, USA

Patrick Shafto
Department of Psychological and Brain Sciences
University of Louisville, Louisville, KY40217, USA

Abstract

Phylogenomic analysis of large sets of molecular characters, primarily DNA and proteins, provides great opportunities to estimate and understand important evolutionary processes. However, molecular phylogenies inferred from individual loci often differ. This incongruence among phylogenies can be the result of systematic error, but can also be the result of different evolutionary histories. We propose a new method, based on Bayesian hierarchical clustering and posterior probability, to measure congruence between genes and to identify sets of congruent loci within which the genes or proteins share identical evolutionary history. We demonstrate the method on a sequence data of 10 nuclear genes from 20 ray-finned fish (*Actinopterygii*) species.

1 Introduction

The availability of genome-scale data provides unprecedented opportunities for phylogenetic analysis (phylogenomics). However, molecular phylogenies inferred from individual loci may conflict with each other (incongruence). The incongruence between genes can be the result of random and systematic errors in phylogenetic tree reconstruction, but can also be caused by the underlying biological processes, including population genetic processes [9], within-species genetic recombination (e.g., chromosomes crossover and gene conversion) [20] and horizontal gene transfer [14].

Techniques for assessing the significance of phylogenetic incongruence are particularly important to systematic biology in a genome-scale. Due to various heterogeneities caused by the biological processes, however, measuring phylogenetic incongruence has been a statistically and computationally challenging task. Nevertheless, several methods have been proposed (Planet [25] provided an excellent review). An intuitive framework for measuring incongruence is the incongruence length difference (ILD) test [5], initially developed in a parsimony context, and later adapted to a distance-based method [32]. The test statistic is defined by $d = L_C - \sum_{i=1}^{N} L_i$ where L_i and L_C denote the lengths of the most parsimonious trees calculated for the ith individual loci and for the combined loci, respectively. However, studies have suggested that the test performs poorly when substantial rate or pattern heterogeneity exists among sites [3, 4].

In a maximum likelihood context, Huelsenbeck and Bull [12] described a method based on likelihood ratio test with the ratio $d = L_1/L_0$ where L_0 is the maximum likelihood assuming all the genes sharing identical trees while allowing rate heterogeneity to vary across sites, and L_1 is the maximum likelihood assuming all the genes undergone different trees and different evolutionary rates. The null distribution for the test is calculated using bootstrapping resampling technique. Based on hierarchical clustering and the likelihood ratio test, Leigh et al. [18] described a method to identify congruent subsets of genes. However, there are several concerns with a maximum-likelihood and bootstrap based approach. To calculate P-values using nonparametric bootstrap, the maximum likelihood estimation must be repeated typically 100 to 1000 times. It therefore can be prohibitively slow [15]. In addition, the empirical test of Hillis and Bull [11] suggested that the bootstrap proportion varied too much among replicate data sets to be used as a measure of repeatability.

Bayesian approach usually models uncertainty in a more interpretable style than maximum likelihood approach. Although Bayesian analysis have been successfully applied to estimate phylogeny, to our knowledge, very few of these works can explicitly test incongruence between genes or identify congruent gene subsets. Most of these analyses assumed that all genes evolved under the same phylogenetic

*corresponding author, eric.rouchka@louisville.edu

tree [13, 15, 16, 24]. Suchard et al. [29] proposed a Bayesian hierarchical model which allowed partitions to have different trees. However, it did not explicitly measure the degree of incongruence among genes. At the same time, it assumed that partitions were known in advance and thus failed in identifying congruent gene subsets. Ané et al. [2] analyzed each gene separately using Bayesian analysis and constructed a gene-to-tree map which is, in turn, used to estimate the posterior probability of pairwise gene dissimilarity. A drawback of this method is that gene trees, exclusively inferred separately, may not resolve well.

Motivated by the shortcomings of existing methods, we propose a Bayesian model to measure incongruence between genes and to identify sets of congruent loci within which genes share identical evolutionary history. From a Bayesian perspective, the method provides a more interpretable and accurate estimation of congruence through the posterior probability of genes being congruent. Based on Bayesian hierarchical clustering [10], the method provides a fast deterministic alternative to Markov Chain Monte Carlo (MCMC) in approximation of the posterior probabilities.

2 Methods

The analysis begins with aligned molecular sequence data Y over N loci, primarily DNA or protein sequences. Data $\mathsf{Y} = (Y_1, \cdots, Y_N)$ consist of N disjoint alignments with Y_n $(n = 1, \cdots, N)$ corresponding to loci n. Data $\mathsf{Y}_k = (Y_{k1}, \ldots, Y_{kN_k})$ denote a subset of Y ($\mathsf{Y}_k \subseteq \mathsf{Y}$) consisting of N_k disjoint alignments, where each Y_{kg} $(g = 1, \ldots, N_k)$ refers to some Y_n $(n = 1, \ldots, N)$.

The first hypothesis, denoted H_0, states that the interesting alignments are congruent. The alternative hypothesis, denoted H_1, states that at least some part of the interesting alignments are incongruent to the others. According to Bayes' theorem, the posterior probability of the N_k markers in Y_k being congruent given the alignment is

$$p(H_0|\mathsf{Y}_k) = \frac{\pi_k p(\mathsf{Y}_k|H_0)}{\pi_k p(\mathsf{Y}_k|H_0) + (1-\pi_k)p(\mathsf{Y}_k|H_1)} \quad (1)$$

The larger $p(H_0|\mathsf{Y}_k)$ is, the more confidence we have in H_0 to believe that the N_k markers are congruent.

The algorithm start with measuring the degree of congruence for all pairs of loci, and the pair with the highest posterior probability (denoted r) is selected. The value r is compared with a threshold p, ($p = .5$ in this work), if $r > p$, the test continues, treating this pair as a congruent gene cluster consisting of two genes. If $r \leq p$, none of the pairs are congruent and the test ends. The algorithm is shown in Figure 1, where congruent information between pair of genes or gene clusters are represented. In Section 2.1, the formal definition of topological congruence and branch length congruence are described. In Section 2.2, a greedy algorithm is proposed to estimate the likelihood quantities involved in the evaluation of the posterior probability defined in Equation 1.

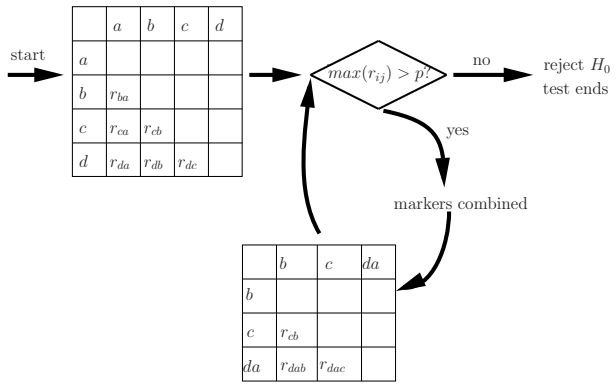


Figure 1: Hierarchical clustering algorithm using posterior probability of gene clusters being congruent as merging criteria. a, b, c, d denote markers.

2.1 Likelihood of Congruence

For an aligned set of sequences $\mathsf{Y}_k = (Y_{k1}, \ldots, Y_{kN_k})$ over N_k loci, topological congruence defines all N_k genes as having identical evolution topology but with various branch lengths and substitution processes. Thus the marginal likelihood that the N_k markers are *topologically congruent* given alignments Y_k is

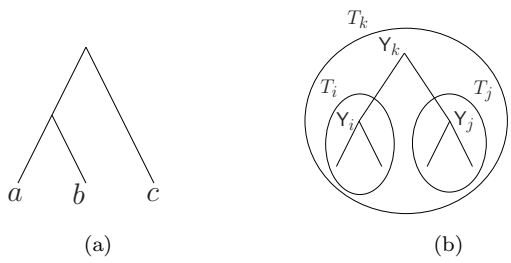
$$p(\mathsf{Y}_k|H_0) = \int \prod_{g=1}^{N_k} p(Y_{kg}|\tau_k, \beta_{kg}, \Theta_{kg}) p(\tau_k, \beta_{kg}, \Theta_{kg}) \, d\,\tau_k, \beta_{kg}, \Theta_{kg} \quad (2)$$

where τ_k is the topology shared by these N_k genes, β_{kg} is the branch length of sequence Y_{kg}, and Θ_{kg} is the substitution model of sequence Y_{kg}.

For branch-length congruence, all N_k genes have identical branch lengths in addition to identical topology, so the marginal likelihood that these N_k loci are *branch-length congruent* given the alignments Y_k is

$$p(\mathsf{Y}_k|H_0) = \int \prod_{g=1}^{N_k} p(Y_{kg}|\tau_k, \beta_k, \Theta_{kg}) p(\tau_k, \beta_k, \Theta_{kg}) \, d\,\tau_k, \beta_k, \Theta_{kg}$$

The rest of the paper focuses on topological congruence. However, the same algorithm can be applied



(a) (b)

Figure 2: (a) An example tree with three genes. Tree-consistent partitions are $a|b|c$ and $ab|c$. (b) A portion of a tree showing T_i and T_j are merged into T_k.

to branch-length congruence with minor modifications. In Section 2.3, the form of the likelihood function given a *single* gene and the strategies on prior distribution are presented. The evaluation of the marginal likelihood (Equation 2) is discussed in Section 2.4.

2.2 Likelihood of Incongruence

A main difficulty when evaluating the marginal likelihood of congruence comes from the hypothesis' combinatorial nature. For example, assume we have three markers (a, b, c). Hypothesis H_1, stating that at least some of the markers are incongruent given the alignments, allows four possibilities: $\{a|b|c,\ ab|c,\ ac|b,\ a|bc\}$, where symbol | separates incongruent markers from congruent markers. Thus the marginal likelihood of H_1 given sequence alignments Y_a, Y_b, Y_c is

$$\begin{aligned} &p(Y_a, Y_b, Y_c|H_1) = w_1 p(Y_a|H_0)p(Y_b|H_0)p(Y_c|H_0) \\ &+ w_2 p(Y_a, Y_b|H_0)p(Y_c|H_0) + w_3 p(Y_a, Y_c|H_0)p(Y_b|H_0) \\ &+ w_4 p(Y_b, Y_c|H_0)p(Y_a|H_0) \end{aligned}$$

where w_i $(i = 1, \ldots, 4)$ is the weight for each case and $\sum_{i=1}^{4} w_i = 1$. A brute force estimation of $p(\mathsf{Y}_k|H_1)$ requires enumerating all possible incongruent clusterings over these N_k markers. Notice that the number of possible clusterings over n elements is the nth bell number: $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$,which prohibits the use of the brute force approach. Instead, we follow the approximation approach developed by Heller and Ghahramani [10] and restrict to clusterings that partition the genes in a manner consistent with the subtrees of the merging algorithm described in Figure 1. For example, if three genes a, b, c are merged according to Figure 2(a), then we only consider two clusterings: $\{a|b|c,\ ab|c\}$. So,

$$\begin{aligned} p(Y_a, Y_b, Y_c|H_1) \approx \big\{&\pi p(Y_a, Y_b|H_0) + \\ &(1-\pi)p(Y_a|H_0)p(Y_b|H_0)\big\}p(Y_c|H_0) \end{aligned}$$

More generally, assume gene cluster k is merged from two mutually exclusively subsets of genes i and j. That is, $\mathsf{Y}_k = \mathsf{Y}_i \cup \mathsf{Y}_j$ and $\mathsf{Y}_i \cap \mathsf{Y}_j = \emptyset$. Equipped with the restricted hypothesis, which we denote $\tilde{H}_1$, the likelihood of incongruence is

$$p(\mathsf{Y}_k|H_1) \approx p(\mathsf{Y}_k|T_k, \tilde{H}_1) = p(\mathsf{Y}_i|T_i)p(\mathsf{Y}_j|T_j) \quad (3)$$

and

$$p(\mathsf{Y}_k|T_k) = \pi_k p(\mathsf{Y}_k|H_0) + (1-\pi_k)p(\mathsf{Y}_k|T_k, \tilde{H}_1) \quad (4)$$

where T_i, T_j, T_k are binary trees expressing the merging processes as shown in Figure 2(b). Restricting to tree-consistent clusterings and assigning different prior probability to them, the method provides a reasonable approximation to the brute force approach which averages over all possible clusterings.

2.3 Likelihood Function and Priors

All sites from an individual gene sequence (e.g., an aligned sequence Y_{kg}) are assumed to evolve under identical topology. Assuming the same substitution rate across sites, however, can be unrealistic. A more nuanced model would allow using one set of substitution parameters for each site. This, however, results in too many parameters to estimate given a limited number of observations. A more practical approach is to model the rate variation using a probabilistic distribution. We use the discrete-gamma model [31].

In the discrete-gamma model, a finite mixture model is used to model across-site rate heterogeneity. All sites within a gene are assumed to share a substitution pattern (based composition or transition-transversion rate), but fall into several classes with different rates. Thus, a site with rate r_c and pattern Q has the substitution-rate matrix $r_c Q$, with r_c calculated using a gamma function. As it is not known to which rate class each site belongs, we average over all the site classes. Incorporating this into the likelihood function, given a sequence alignment Y_{kg} of gene kg, we have

$$\begin{aligned} &p(Y_{kg}|\tau_k, \beta_{kg}, \Theta_{kg}) \\ &= \prod_{s=1}^{S_{kg}} \sum_{c=1}^{C} p(Y_{kgs}|\tau_k, \beta_{kg}, r_c Q_{kg}) p(r_c) \end{aligned} \quad (5)$$

where Y_{kgs} denotes the sth site in sequence Y_{kg}, S_{kg} is the number of sites in Y_{kg}, and Q_{kg} is the substitution pattern shared by all sites within Y_{kg}. The summation is a weighted average over all C site-rate classes. $p(r_c)$ is the prior probability that a site's rate falls in rate class c. For equally likely rate classes, $p(r_c) = 1/C$.

For the general time-reversible (GTR) model of nucleotide substitution, the matrix is normally written as the product of a symmetric matrix R representing

substitution rate, and a diagonal matrix Π representing a stationary distribution:

$$Q_{kg}^{\mathrm{GTR}} = R_{kg}\Pi_{kg} = \begin{pmatrix} \cdot & a_{kg}\pi_{kgC} & b_{kg}\pi_{kgA} & c_{kg}\pi_{kgG} \\ a_{kg}\pi_{kgT} & \cdot & d_{kg}\pi_{kgA} & e_{kg}\pi_{kgsG} \\ a_{kg}\pi_{kgT} & d_{kg}\pi_{kgC} & \cdot & f_{kg}\pi_{kgG} \\ c_{kg}\pi_{kgT} & e_{kg}\pi_{kgC} & f_{kg}\pi_{kgA} & \cdot \end{pmatrix}$$

Once the tree topology, branch lengths, and site-specific rates are chosen, the likelihood at each site ($p(Y_{kgs}|\tau_k, \beta_{kg}, r_c Q_{kg})$) and the likelihood for each gene (see Equation 5) are computed using Felsenstein's pruning algorithm [6].

The stationary distribution requires summation to one and so is modeled by a Dirichlet prior distribution,

$$\mathrm{diag}(\Pi_{kg}) \sim \mathrm{Dirichlet}(\alpha_{kg}).$$

The tree topology is sampled from a multinomial distribution,

$$\tau_k \sim \mathrm{Multinomial}(p_1, \ldots, p_E).$$

where $E = (2M-5)!/2^{M-3}(M-3)!$, p_i $(i = 1, \ldots, E)$ is the probability of the ith topology being sampled over the E possible M-taxon topologies. Without bias, these E topologies are assumed to be equally probable, so $p_i = 1$ $(i = 1, \ldots, E)$.

The prior information for branch lengths within a gene is modeled by an exponential distribution with an average branch length $1/\lambda_{kg}$,

$$\beta_{kg} \sim \mathrm{Exponential}(1/\lambda_{kg}).$$

The prior belief on a set of genes being congruent is expressed using π_k (as in Equation 1). $\pi_k = 0$ expresses a strong belief that alignments in Y_k are incongruent, while $\pi_k = 1$ says they are congruent. The Dirichlet process prior [1] is used to model the prior belief. Assume a set of genes are partitioned into congruent gene clusters of various sizes (here size means the number of genes in a cluster). For a new gene not in this set, a Dirichlet process prior, in general, says that this new gene is more likely to be congruent with gene clusters of larger size. Heller and Ghahramani [10] proposed a prior for agglomerative clustering, which has similar property to Dirichlet Process prior:

$$\begin{array}{lll} \pi_k = 1 & d_k = \eta & \text{if } T_k \text{ is a leaf node} \\ \pi_k = \frac{\eta\Gamma(N_k)}{d_k} & d_k = \eta\Gamma(N_k) + d_i d_j & \text{else} \end{array}$$

where η is the concentration hyperparameter.

In this work, $\alpha_{kg} = (1,1,1,1)$, $\lambda_{kg} = 10$ for all k and g, and $\eta = 0.5$, though a Bayesian hierarchical model can be easily built such that the uncertainty on hyperparameters α_{kg}, λ_{kg}, and η are incorporated into the model.

2.4 Estimation of Marginal Likelihood

A key computation component of the model described in Section 2.1 is the calculation of the marginal likelihood defined in Equation 2, which is a highly variable function over a high dimensional parameter space. The integral is analytically intractable (e.g. due to lack of conjugate priors), and the parameter space is too high-dimensional for numerical integration. In this work, the approach by Newton and Raftery [22] using Monte Carlo sample from the posterior is used. Notice that marginal likelihood can be expressed as an expectation with respect to the posterior distribution of the parameters:

$$\frac{1}{p(\mathsf{Y}_k|H_0)} = \int \frac{p(\Omega_k|\mathsf{Y}_k)}{p(\mathsf{Y}_k|\Omega_k)} d\Omega_k = E\left\{ \frac{1}{p(\mathsf{Y}_k|\Omega_k)} \middle| \mathsf{Y}_k \right\} \tag{6}$$

where $\Omega_k = (\tau_k, \beta_{kg}, \Theta_{kg}), g = 1, \ldots, N_k$ are model parameters, and $p(\mathsf{Y}_k|\Omega_k)$ are the likelihood function, as indicated in Equation 2. From here the harmonic mean identity can be used to approximate the marginal likelihood $p(\mathsf{Y}_k|H_0)$:

$$\hat{p}(\mathsf{Y}_k|H_0) = \left\{ \frac{1}{S} \sum_{t=1}^{S} \frac{1}{p(\mathsf{Y}_k|\Omega_k^t)} \right\}^{-1} \tag{7}$$

where $\Omega_k^1, \ldots, \Omega_k^S$ are S samples drawn from the posterior distribution $p(\Omega_k|\mathsf{Y}_k)$.

MCMC has been widely used in phylogenetic inference to sample model parameters [13, 15, 28]. The approach in MRBAYES [13] is adapted in this work. To draw from $p(\Omega_k|\mathsf{Y}_k)$, the sampler uses a Metropolis-within-Gibbs [30] algorithm that cycles through blocks of model parameters within Ω_k, updating them via a Metropolis-Hastings proposal. For example, to sample the substitution model parameters for the first markers in Y_k, the acceptance probability is:

$$r = min\left(1, \frac{p(\Theta_{k1}^*)}{p(\Theta_{k1})} \frac{p(\mathsf{Y}_{k1}|\tau_k, \beta_{k1}, \Theta_{k1}^*)}{p(\mathsf{Y}_{k1}|\tau_k, \beta_{k1}, \Theta_{k1})} \frac{q(\Theta_{k1}|\Theta_{k1}^*)}{q(\Theta_{k1}^*|\Theta_{k1})}\right) \tag{8}$$

where Θ_{k1}^* stands for the proposed values for the substitution model parameters. Simulated tempering [21], also known as Metropolis-coupled MCMC [8], is used to reduce the chance that Markov chain simulations remain in the neighborhood of a single model for a long period of time.

It is worth noting that estimation of marginal likelihood remains a central problem in Bayesian inference. The decision of using the harmonic mean estimator is due to its simplicity. However, the estimator can have infinite variance. Raftery et al. [26] described a stabilized version of the estimator. Gelman and

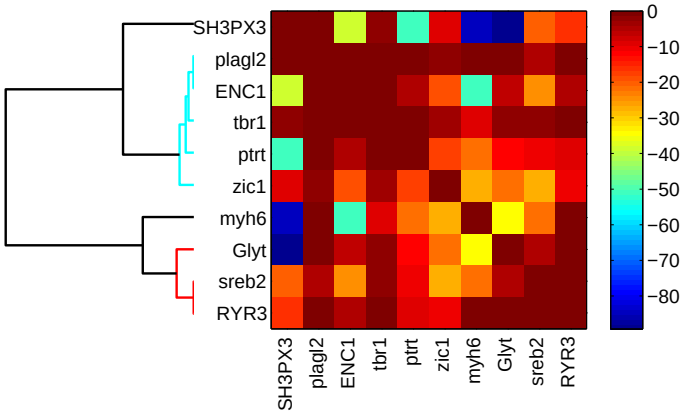


Figure 3: The dendrogram shows the hierarchical clustering structure of genes based on their posterior probability of being congruent. The square heatmap shows the congruence relationships between pairs of genes. The warmer the color is in a cell, the more congruent the corresponding pair of genes are. The colormap shows values of posterior probability (in logarithm) represented by colors.

Meng [7] proposed path sampling which generalizes the thermodynamic integration originated from theoretical physics and involves a sequence of intermediate distributions bridging prior and posterior. Lartillot and Philippe [17] applied thermodynamic integration to phylogenetic analysis.

3 Results

The method proposed herein is used to estimate the phylogeny relationships amongst ray-finned fish (*Actinopterygii*) with 10 alignments of protein-coding genes assembled by Li et al. [19]. Twenty species, out of 52 ray-finned fish, are randomly selected, and mouse (*Mus musculus*) is used as the outgroup to root the phylogeny tree. Li et al. [19] defined one data block for each codon position and each gene, yielding 30 data blocks (3 codon positions $\times$ 10 genes). For each data block, substitution parameters (GTR + Γ) were estimated using maximum likelihood and Bayesian inference method. They defined the distance between data blocks using their estimated substitution parameters. Then data blocks were clustered by hierarchical clustering with centroid linkage. As expected, the three major clusters discovered by their method corresponded exactly to codon positions. The trees inferred from each individual gene by the Bayesian phylogenetic method (MRBAYES GTR + Γ) either are poorly resolved star-like trees or exhibit obviously different topology (data not shown here), indicating that a systematic way of combining these genes is desirable in order to accurately analyze the data set.

The Bayesian topological congruence method proposed herein is applied to identify congruent sets of genes using a Dirichlet process prior with concentration parameter $\eta = .5$. From this test, four mutually incongruent sets of genes were identified, containing 5, 3, 1, and 1 genes, respectively. The pairwise gene congruence is shown in a square matrix in Figure 3. The warmer (e.g., red is warmer than blue) the color is in a cell, the more congruent the corresponding pair of genes are. The colorbar maps color to values of posterior probability (on a logarithmic scale). The degree of congruence between genes ranges from extremely congruent to extremely incongruent. Gene pairs such as (*plagl2, ENC1*), (*tbr1, ptrt*) are very congruent, with posterior probabilities near 1; gene pairs such as (*myh6, SH3PX3*), (*ENC1, myh6*) are very incongruent, with posterior probabilities smaller than e^{-50}. It also indicates that some genes, such as *SH3PX3* and *myh6* are incongruent to most of the other genes.

Genes are further clustered into congruent subsets, shown in a dendrogram in Figure 3. Branch lengths in the dendrogram correspond to the posterior probability of congruence between gene subsets connected by the branch. The shorter the branch, the more congruent they are. The cut point value is $p = 0.5$. Branches having $r \leq 0.5$ are in black and $r > 0.5$ are in lighter colors. The tree shows two main congruent subsets: set1=(*plagl2, ENC1, tbr1, ptrt, zic1*) and set2=(*RYR3, sreb2, Glyt*). Notice that although *SH3PX3* are congruent to *plagl2* and to *tbr1*, it is not included in congruent set 1 since that merge has the posterior probability $r = e^{-67}$. This is also indicated in the square matrix, where the first column shows that *SH3PX3* is incongruent to *ENC1* and *ptrt*. Similarly, although *myh6* is highly congruent with *RYR3*, the gene is not included in congruent set 2 because the merge has posterior probability $r = e^{-50}$.

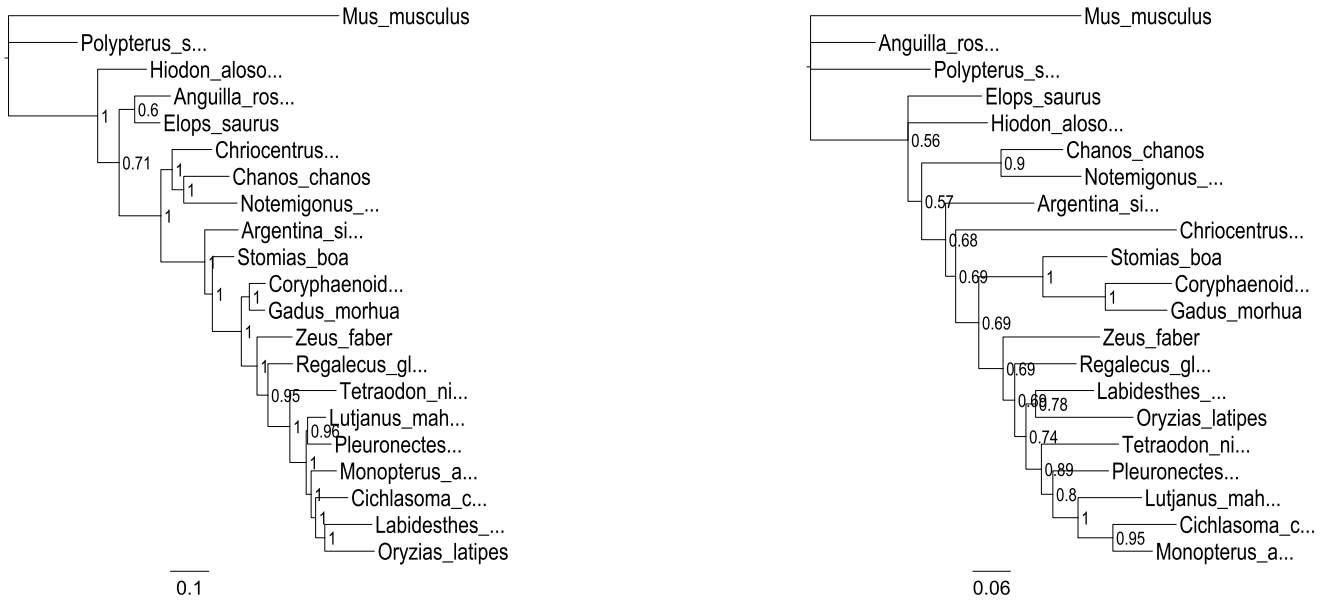


Figure 4: 50% majority-rule consensus trees inferred from congruent set 1 (a) and congruent set 2 (b). Posterior probabilities for branches are indicated.

Bayesian phylogenies inferred from each of the two congruent sets are shown in Fig. 4. Branches of the 50% majority-rule consensus tree from congruent set 1 have high posterior probability, providing strong support for the topology. The main branch with low probability is the pair (*Anguilla_rostrata, Elops_saurus*). Although the 50% majority-rule consensus tree from congruent set 2 has an overall similar topology as the one from congruent set 1, its branches have relatively low posterior probabilities. However, one interesting result comes from analysis of congruent set 2. In this set, there are three levels of ancestral nodes from the *Chriocentrus_dorab* group to the (*Chanos_chanos, Notemigonus_crysoleucas*) group, while in congruent set 1, *Chriocentrus_dorab* and the (*Chanos_chanos, Notemigonus_crysoleucas*) group share an immediate common ancestor.

4 Discussion

Bayesian methods of multigene analysis correspond to various ways of partitioning the genome [16, 23, 24, 27]. Gene topological congruence analysis can be considered as partitioning genes according to the underlying gene topology while allowing branch length and substitution heterogeneity within a partition. To infer gene partitioning based on topological congruence, a mixture model is proposed:

$$p(\mathsf{z}|\mathsf{Y}) = \frac{p(\mathsf{Y}|\mathsf{z})p(\mathsf{z})}{p(\mathsf{Y})} = \frac{\prod_{k=1}^{K} L(\mathsf{Y}_k|H_0)p(\mathsf{z})}{p(\mathsf{Y})} \qquad (9)$$

where, if N genes are clustered into K partitions, $\mathsf{z} = (z_1, \ldots, z_N)$, z_i is the partition of the ith gene, and $L(\mathsf{Y}_k|H_0)$ is the marginal likelihood integrating over heterogeneous parameters, as defined by Equation 2 for topological congruence.

The posterior probability of multiple markers (for example, three markers: a, b, c) being congruent given the sequences are the posterior probability of them being assigned into one partition:

$$p(H_0|\mathsf{Y}) = \sum_{\substack{\mathsf{z} \in \mathcal{Z} \\ z_a = z_b = z_c}} p(\mathsf{z}|\mathsf{Y}) \qquad (10)$$

where $\mathcal{Z}$ is the set of all possible clusterings over N elements. Although MCMC inference algorithm has been widely used for phylogenetic analysis, sampling over the large sample space imposed by Equation 9 is extremely computationally expensive.

The greedy agglomerative algorithm in Figure 1 can be considered as a deterministic alternative to estimating the mixture model (Equation 9) by a sampling method such as MCMC [10]. However, it must be noted that this method still does not scale well with very large numbers of loci for two reasons. First, the

agglomerative algorithm (Figure 1) has a computation time complexity of $O(N^2)$, where N is the number of genes in the data set. Second, the merging criterion still requires calculating the marginal likelihood (Equation 2) using an MCMC sampler. For this reason, the experiment reported in this work includes only ten genes and twenty taxa, a data set smaller than would be normally interesting to genome wide phylogenetic analysis.

In general, Bayesian phylogenomic analysis methods that account for evolutionary heterogeneity among genes, including the algorithm described in this work, can present significant computational challenges. One solution is to devise parallelizable algorithms. It is particularly interesting to point out that the algorithm presented in this work is readily parallelizable. For example, given three gene clusters i, j and k, the evaluation of $p(H_0|\mathsf{Y}_i, \mathsf{Y}_j)$ and $p(H_0|\mathsf{Y}_i, \mathsf{Y}_k)$ are independent and can therefore be computed in parallel by different machines. This can significantly speed up the computation and allow much larger scale applications of the algorithm.

5 Conclusion

Genomic scale data offers invaluable opportunities to solve difficult phylogenetic problems, but also imposes enormous challenges for statistical and computational methods [27]. The method proposed in this work accounts for evolutionary heterogeneities and identifies congruent gene subsets using Bayesian hypothesis testing. The proposed method approximates the posterior probability of genes being congruent in a fast deterministic manner. A notable feature of the method is that it is particularly suitable for parallel computation. The test presented on the data set shows that the model recovers interesting congruence structure among genes. Future work will explore applications of the model to more interesting genome wide data.

Acknowledgements

We would like to thank Dr. Ming Ouyang for his advice on the topic and comment on the manuscript.

References

[1] David Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII – 1983*, pages 1–198, 1985.

[2] Cécile Ané, Bret Larget, David A. Baum, Stacey D. Smith, and Antonis Rokas. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, 24(2):412–426, 2007.

[3] Pierre Darlu and Guillaume Lecointre. When does the incongruence length difference test fail? *Mol Biol Evol*, 19 (4):432–437, 2002.

[4] Konrad Dolphin, Robert Belshaw, Orme, and Donald L. Quicke. Noise and incongruence: Interpreting results of the incongruence length difference test. *Molecular Phylogenetics and Evolution*, 17(3):401–406, 2000.

[5] James S Farris, Mari Kallersjo, Arnold G. Kluge, and Carol Bult. Testing significance of incongruence. *Cladistics*, 10 (3):315–319, 1994.

[6] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[7] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 1998.

[8] Charles J. Geyer. Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, 1991.

[9] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.

[10] Katherine A. Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 297–304. ACM Press, 2005.

[11] David M. Hillis and James J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology*, 42(2):182–192, 1993.

[12] John P. Huelsenbeck and J. J. Bull. A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, 45(1):92–98, 1996.

[13] John P. Huelsenbeck and Fredrik Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[14] Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96:3801–3806, 1999.

[15] Bret Larget and Donald L Simon. Markov chasin monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol*, 16:750–759, 1999.

[16] Nicolas Lartillot and Herve Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6):1095–1109, 2004.

[17] Nicolas Lartillot and Herve Philippe. Computing bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

[18] Jessica W. Leigh, Edward Susko, Manuela Baumgartner, and Andrew J. Roger. Testing congruence in phylogenomic analysis. *Systematic Biology*, 57(1):104–115, 2008.

[19] Chenhong Li, Guoqing Lu, and Guillermo Orti. Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. *Systematic Biology*, 57(4):519–539, 2008.

[20] Matthew S. Meselson and Charles M. Radding. A general model for genetic recombination. *Proceedings of the National Academy of Sciences*, 72(1):358–361, 1975.

[21] Radford Neal. Sampling from multimodal distributions using tempered transitions. *Journal Statistics and Computing*, 6(4):353–366, 1996.

[22] Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap.

J. Roy. Statist. Soc., (B 56):3–48, 1994.

[23] Johan A.A. Nylander, Fredrik Ronquist, John P. Huelsenbeck, and Joséluis Nieves-Aldrey. Bayesian phylogenetic analysis of combined data. *Systematic biology*, 53(1):47–67, 2004.

[24] Mark Pagel and Andrew Meade. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol*, 53(4):571–581, 2004.

[25] Paul J. Planet. Tree disagreement: Measuring and testing incongruence in phylogenies. *Journal of Biomedical Informatics*, 39(1):86–102, 2006.

[26] Adrian E. Raftery, Michael A. Newton, Jaya M. Satagopan, and Pavel N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics 8*, pages 1–45, 2007.

[27] Bruce Rannala and Ziheng Yang. Phylogenetic inference using whole genomes. *Annual review of genomics and human genetics*, 9(1):217–231, 2008.

[28] Marc A. Suchard, Robert E. Weiss, and Janet S. Sinsheimer. Bayesian selection of continuous-time markov chain evolutionary models. *Mol Biol Evol*, 18(6):1001–1013, 2001.

[29] Marc A. Suchard, Janet S. Kitchen, Christina M.R.and Sinsheimer, and Robert E. Weiss. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology*, 52(5):649–664, 2003.

[30] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

[31] Ziheng Yang. *Computational Molecular Evolution (Oxford Series in Ecology and Evolution)*. Oxford University Press, USA, 2006.

[32] Marina Zelwer and Vincent Daubin. Detecting phylogenetic incongruence using bionj: an improvement of the ild test. *Molecular phylogenetics and evolution*, 33(3):687–693, 2004.