An Application of Random Walk Resampling to Phylogenetic HMM Inference and Learning

Wei Wang
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
wangwe90@msu.edu

Qiqige Wuyun
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
wuyunqiq@msu.edu

Kevin J. Liu
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
kjl@msu.edu

***Abstract*—Statistical resampling methods are widely used for confidence interval placement and as a data perturbation technique for statistical inference and learning. An important assumption of popular resampling methods such as the standard bootstrap is that input observations are identically and independently distributed (i.i.d.). However, within the area of computational biology and bioinformatics, many different factors can contribute to intra-sequence dependence, such as recombination and other evolutionary processes governing sequence evolution. The SEquential RESampling ("SERES") framework was previously proposed to relax the simplifying assumption of i.i.d. input observations. SERES resampling takes the form of random walks on an input of either aligned or unaligned biomolecular sequences.**

This study introduces the first application of SERES random walks on aligned sequence inputs and is also the first to demonstrate the utility of SERES as a data perturbation technique to yield improved statistical estimates. We focus on the classical problem of recombination-aware local genealogical inference. We show in a simulation study that coupling SERES resampling and re-estimation with recHMM, a hidden Markov model-based method, produces local genealogical inferences with consistent and often large improvements in terms of topological accuracy. We further evaluate method performance using an empirical HIV genome sequence dataset.

***Index Terms*—phylogenetic, hidden Markov model, SERES, random walk, biomolecular sequence, recombination, gene tree, genealogy, statistical inference and learning, HIV**

I. Background

Statistical resampling methods are widely used in science and engineering. Among the many applications of resampling methods is calculating confidence intervals for statistical inference and learning [Efron, 1979, Efron and Tibshirani, 1986]. Another important application arises in the context of statistical inference and learning. Alongside model perturbation approaches such as dropout [Srivastava et al., 2014], statistical resampling can be seen as a form of data perturbation that can help to improve inference and learning accuracy [Breiman, 1996].

Two classes of resampling methods are used. The first class is non-parametric; of these, the bootstrap method is among the most widely used [Efron, 1979, Efron and Tibshirani, 1986]. Given an input set of observations, the bootstrap method resamples observations uniformly at random with replacement. Re-estimation is then performed on resampled replicates, and repeatability is quantified by comparing re-estimates. Other related non-parametric methods include the jackknife, weighted bootstrap, and others. In contrast, parametric methods resample directly from an explicit statistical model. Ideally, the model that generated the original inputs is available, but in practice a hypothesis model must typically be assumed. Non-parametric methods are a popular choice since they avoid the need to assume that observations were generated under a specific model.

But the bootstrap method and other popular non-parametric resampling methods bring their own important limitation: the simplifying assumption that input observations are independent and identically distributed (i.i.d.). The i.i.d. assumption is invalid in the case where inputs consist of sequences of observations, as is common throughout genomics and many other topics in computational biology and bioinformatics.

To relax this simplifying assumption, we developed the SERES (or "SEquential RESampling") method for non-parametric/semi-parametric resampling from an input of either aligned or unaligned sequences [Wang et al., 2018]. SERES synthesizes and extends the bootstrap method with a simple but powerful insight due to Landan and Graur [2007]: inferences should be repeatable whether an input of unaligned sequences is read left-to-right or right-to-left. In lieu of using "mirrored" inputs, SERES performs random walks on input sequences. In this study, we focus on the SERES algorithm for aligned sequence inputs. A start point (i.e., an initial site) and direction for the SERES walk are chosen uniformly at random. The walk then proceeds, where aligned sites are sampled during each

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

step of the walk; walk reversals occur with probability γ during each step (where γ is typically smaller than 0.5) and with certainty at the start and end of the alignment. The random walk concludes once the resampled replicate has length equal to the input alignment. For each resampled replicate, re-estimation is performed. Repeatability is then measured by quantifying disagreement among re-estimations.

Our initial study of SERES focused on the SERES algorithm for unaligned sequence inputs [Wang et al., 2018], rather than aligned inputs. Briefly, the SERES algorithm for unaligned sequence inputs also takes the form of a random walk, with one main difference: resampling "reads" along unaligned sequences occur in an asynchronous fashion, and a set of anchors serve as synchronization "barriers" in much the same sense as in parallel computing. We previously applied the SERES algorithm to perform confidence interval placement for a classical problem in computational biology and bioinformatics: multiple sequence alignment (MSA) estimation. Using synthetic and empirical data, we showed that the use of SERES random walks within a resampling/re-estimation pipeline resulted in comparable or often better type I and type II error rates relative to state-of-the-art methods.

In this study, we address several corollary questions which constitute the three primary contributions of our study. (1) We propose the first application of SERES random walks on aligned sequences, whereas our earlier study focused on SERES random walks on unaligned sequences. (2) Our study utilizes SERES random walks as a means to "boost" HMM inference/learning performance. Like other non-parametric resampling methods, we show that SERES has utility as a data perturbation technique in addition to its use in confidence interval placement, as considered by our earlier work. (3) We introduce a SERES-based approach for another classical problem in computational biology and bioinformatics: recombination-aware local genealogical inference.

II. Methods

Standalone recHMM analysis

The coalescent-with-recombination (CwR) model [Hudson, 1983] is a classical population genetic model involving recombination. However, phylogenetic inference under the multi-species CwR model is computationally prohibitive, and alternatives such as the sequentially Markovian coalescent (SMC) model [McVean and Cardin, 2005] are used as an approximation to the full CwR model. First-order hidden Markov models (HMMs) are a widely used choice for tractable SMC-based inference.

Phylogenetic HMMs (or "phylo-HMMs") are the class of HMMs with hidden states that correspond to phylogenies. Markovian dependence between phylo-HMM states are meant to capture intra-sequence dependence among local phylogenies, which can be caused by recombination and other evolutionary processes. There are a variety of phylo-HMM-based methods for local genealogical inference, depending on modeling assumptions [Husmeier and Wright, 2001, Liu et al., 2014, Mailund et al., 2011, Westesson and Holmes, 2009]. We focus on recHMM [Westesson and Holmes, 2009] as an exemplar method in this class.

The recHMM framework utilizes a statistical model that combines a finite-sites substitution model and a phylo-HMM to capture intra-sequence dependence due to recombination. The combined model parameters θ consist of local gene tree branch lengths, substitution model rates and base frequencies, and state transition probabilities. Emissions occur at a state with likelihood under a finite-sites substitution model, which can be efficiently calculated using Felsenstein's peeling algorithm [Felsenstein, 2004]. Combined model likelihood is calculated using dynamic programming in the form of either the forward or backward algorithm [Rabiner, 1989]. Typically, model parameters in a traditional HMM are learned by addressing computationally difficult optimization problems; for this reason, heuristics such as the expectation-maximization (EM) algorithm and the related Baum-Welch algorithm are used. An EM-based approach is used to learn recHMM model parameters θ. Westesson and Holmes [2009] also applied a structural EM heuristic [Friedman et al., 2002] to automatically learn the set of local gene trees represented by recHMM's states – one distinct gene tree per state. The recHMM framework allows the user to specify the HMM state space size ϕ. In our simulation study, recHMM was run with a default setting of $\phi = 3$; we also included model reduction experiments with alternative settings $\phi \in \{10, 15\}$. We note that, in the structural EM used by Westesson and Holmes [2009], HMM states are distinguished by both gene tree topologies and branch lengths.

As is common practice for local search heuristics in the context of statistical learning, Westesson and Holmes [2009] utilized ψ independent optimization trials and then selected the best trial under the maximum likelihood criterion as a means to avoid getting stuck in local optima (cf. Figure 1 in [Westesson and Holmes, 2009]). We followed their practice in our study: when run as a standalone method, recHMM utilized $\psi = 100$ independent optimization trials.

Consistent with the study of Westesson and Holmes [2009], we used the posterior decoding algorithm to perform statistical inference of local phylogenies [Rabiner, 1989]. The posterior decoding algorithm addresses the following problem. Let G be the set of all possible unrooted tree topologies on n taxa. The input consists of a multiple sequence alignment A on n sequences – one for each of n taxa – with length k (i.e., k sites in A). A is assumed to contain recombinant sequences, and historical recombination can cause local genealogies to vary across the sites in A [Hein et al., 2004]. The output consists of the following: for each aligned site a_i where $1 \leq i \leq k$, we seek the conditional probability that the HMM is in a hidden state corresponding to a particular gene tree $g \in G$ conditional on all sites in A and the fitted HMM model. For a particular HMM instance, the posterior decoding effectively estimates which gene tree is the most likely evolutionary history that explains the observed character at a given site conditional on the sequence of all observed sites in A. Analogous to the distinction between filtering and smoothing [Russell and

Norvig, 2016], the posterior decoding weighs any particular inference at a given site against the total evidence across all sites.

The SERES+recHMM pipeline

The key algorithmic contribution of this study takes the form of a methodological pipeline for local phylogenetic inference which augments recHMM with SERES random walks. First, we ran SERES resampling on the input alignment A. The Appendix includes detailed pseudocode for this procedure (see Algorithm 1 in Appendix, which is reproduced from [Wang et al., 2018]) as well as an illustrated example of a SERES random walk on an input MSA (Supplementary Figure S1 in Appendix). The SERES resampling procedure in our simulation study utilized a default reversal probability $\gamma = 0.005$. We also conducted additional experiments with alternative reversal probability values $\gamma \in \{0, 0.01, 0.1\}$. The resampling procedure generated 10 SERES replicates per dataset in our study.

Next, we ran recHMM on each SERES replicate. Consistent with the study of Westesson and Holmes [2009], we observed that the quality of recHMM's inference depends upon sufficiently intensive learning optimization. We adopted a conservative approach and restricted the number of independent learning trials ψ used in the SERES-based pipeline, where recHMM was run on each SERES replicate with $\psi = 10$ independent trials. For each dataset, the total number of independent learning trials used in the SERES-based pipeline was therefore equal to the number of independent learning trials used by the standalone recHMM method. Otherwise, recHMM re-estimation of a SERES replicate was run in an identical manner compared to the standalone recHMM method.

Given optimized model parameter values, inference proceeded via the posterior decoding algorithm. The resulting output annotation consists of a per-site probability distribution over ϕ gene tree topologies.

For each site, inferred posterior decoding probability distributions were aggregated across all SERES replicates in which the site appeared (with per-replicate multiplicity based on the number of times that the site was sampled within the replicate). The aggregated distribution was then normalized to obtain a valid probability distribution.

Simulated datasets

Gene trees were simulated under the CwR model using ms [Hudson, 2002]. Each CwR simulation sampled either 4, 5, or 6 alleles with scaled recombination rate $\rho \in \{0.5, 1.0, 2.0\}$ and total sequence length of 1 kb per replicate. For each gene tree, finite-length sequence evolution was simulated under the Jukes-Cantor model of nucleotide substitution [Jukes and Cantor, 1969] using Seq-Gen [Rambaut and Grassly, 1997]. We used a substitution rate $\theta \in \{0.5, 1.0, 2.0\}$. A model condition consisted of fixed values for all model parameters, and simulation procedures were repeated so that 30 replicate datasets were generated per model condition. Summary statistics for the simulated datasets are shown in Supplementary Table S1 in the Appendix. We assessed topological accuracy of inferred gene trees relative to ground truth using the Robinson-Foulds measure [Robinson and Foulds, 1981], which is the proportion of bipartitions that occur in an inferred gene tree but not the true gene tree or vice versa.

The coalescent simulations were performed using the following ms [Hudson, 2002] command:

```
ms <number of taxa>
   1
   -r <rho>
   <number of sites> -T
```

where the number of taxa is 4, 5, or 6, the scaled recombination rate ρ is 0.5, 1, or 2, the number of sites is 1000, and the -T parameter outputs true local gene trees. The Seq-Gen simulations made use of the following command:

```
seq-gen -mHKY -l <number of sites>
        -p <number of partitions>
        -s <mutation rate>
        -z <PRNG seed>
        <gene trees> > <seqfile>
```

where the -mHKY parameter specifies the Jukes-Cantor mutation model, the -l parameter specifies sequence length of 1000 bp, the -p parameter is the number of local gene trees, the -s parameter specifies the mutation rate $\theta \in \{0.5, 1, 2\}$, the -z option specifies the pseudo-random number generator seed, and the <gene trees> argument is the list of true gene trees that were output by ms.

Empirical datasets

We also re-analyzed an HIV dataset from the study of Westesson and Holmes [2009]. The dataset consisted of Indian samples that were originally studied by Lole et al. [1999]. The dataset was subsampled to include four sequences, including the putatively recombinant sequence that was the original foci of the two studies.

Software and data availability

Open-source software and open data can be found at https://gitlab.msu.edu/liulab/seres-based-recombination-breakpoint-inference-data-and-scripts.

III. Results and Discussion

Simulation study

The performance measures evaluated the extent to which each method's inferred per-site posterior probability distribution reflected topological accuracy. We initially examined the correlation between each method's inferred per-site posterior probability for a gene tree topology g and the topological accuracy of g. Equivalently, we quantified the anticorrelation between the former and the topological error of g, as measured by the Robinson-Foulds distance between g and the true gene tree topology for a site. We focus on correlation rather than anticorrelation to simplify discussion. Table I shows correlation results for the 4-taxon model conditions. Across

all 4-taxon model conditions, SERES+recHMM inference was consistently better correlated with topological accuracy compared to standalone recHMM. Performance improvement obtained by coupling recHMM analysis with SERES resampling and re-estimation was robust to varying mutation rates and recombination rates. Absolute correlation improvements were large in magnitude – amounting to at least 0.203 for any model condition and as much as 0.305.

We next compared the two methods' inferred posterior probability distributions on the 4-taxon model conditions. Figure 1 shows a histogram of recHMM-inferred per-site posterior probabilities for gene tree topologies falling into two classes: either the true gene tree topology for a site ("true class") or all other topologies ("false class"); Figure 2 shows the equivalent histogram for SERES+recHMM. Focusing on the true class of per-site inferences, inferences with less than 10% posterior probability were consistently reduced across all model conditions when comparing standalone recHMM versus SERES+recHMM; the reduction amounted to more than half in all cases. The latter method's posterior probability distribution was shifted rightward compared to the former method (i.e., the SERES+recHMM-inferred posterior probability mass was instead distributed among per-site inferences with higher posterior probability, relative to standalone recHMM). The effect was most pronounced for per-site inferences in the highest decile range of posterior probability (i.e., 90% posterior probability or greater). An opposite trend was observed for the false class of per-site inferences. Standalone recHMM's per-site inferences in the highest decile range of posterior probability (i.e., 90% posterior probability or greater) were typically the second highest in frequency compared to all other deciles; in contrast, SERES+recHMM consistently returned fewer per-site inferences in the top decile of posterior probability range – at most a few percentage points and nearing zero frequency for some model conditions. The SERES+recHMM-inferred posterior distribution for the false class of per-site inferences was consistently shifted leftward compared to standalone recHMM. We attribute these findings to two factors. First, the use of non-parametric resampling and re-estimation appears to be conducive to improved inference of true gene tree topologies. Second, incorrect inferences for all other gene tree topologies (in terms of relatively high inferred posterior probability) were less repeatable.

A similar performance outcome was observed on the larger 5-taxon model conditions. Across all model conditions, SERES+recHMM inference was more strongly correlated with topological accuracy compared to recHMM (Table II). We observed absolute improvements in correlation coefficients amounting to between 0.217 and 0.345. Taken together, SERES+recHMM's performance advantage relative to standalone recHMM was larger on the 5-taxon model conditions, relative to the smaller 4-taxon model conditions.

As in the 4-taxon and 5-taxon dataset comparisons, SERES+recHMM's per-site inference was more strongly correlated with topological accuracy across all 6-taxon model conditions, when compared to standalone recHMM (Table III). However, the observed correlation coefficients for both methods were generally weaker when comparing the 6-taxon dataset analyses versus analyses of smaller datasets; furthermore, the observed improvement in correlation returned by SERES+recHMM versus standalone recHMM was smaller as well – ranging from nearly comparable (an absolute improvement of 0.004) to at most 0.206. The histogram comparison of each method's per-site inferences was also different for the true class of per-site inferences, but not for the false class (Supplementary Figures S2 and S3). The latter was in fact consistent: for the false class of per-site inferences, SERES+recHMM's inferred posterior probability distributions were more strongly shifted leftward compared to recHMM. The effect was preserved even though posterior probabilities of the false class of inferences was more than double that seen on the 4-taxon and 5-taxon experiments. However, a different outcome was observed for the true class of per-site inferences: rather than a rightward shift, SERES+recHMM returned posterior probability distributions which were generally more diffuse compared to recHMM alone. We attribute these findings to the increased computational complexity of HMM

TABLE I

On 4-taxon model conditions, posterior probabilities inferred using SERES+recHMM were more highly correlated with topological accuracy compared to standalone recHMM. FOR EACH METHOD'S INFERENCE, WE CALCULATED THE PEARSON CORRELATION BETWEEN THE INFERRED POSTERIOR PROBABILITY FOR A GENE TREE g AND THE TOPOLOGICAL DISTANCE BETWEEN g AND THE TRUE EVOLUTIONARY HISTORY OF A SITE (I.E., THE TRUE LOCAL GENE TREE). AVERAGE CORRELATION FOR A METHOD IS REPORTED ACROSS ALL REPLICATES IN A MODEL CONDITION ($n = 30$).

Number of taxa	Recombination rate ρ	Mutation rate θ	recHMM correlation	SERES+recHMM correlation
4	0.5	0.5	-0.547	-0.830
4	0.5	1	-0.622	-0.866
4	0.5	2	-0.554	-0.799
4	1	0.5	-0.470	-0.775
4	1	1	-0.460	-0.742
4	1	2	-0.427	-0.677
4	2	0.5	-0.560	-0.855
4	2	1	-0.664	-0.867
4	2	2	-0.609	-0.837

TABLE II

On 5-taxon model conditions, posterior probabilities inferred using SERES+recHMM were more highly correlated with topological accuracy compared to standalone recHMM. OTHERWISE, TABLE LAYOUT AND DESCRIPTION ARE IDENTICAL TO TABLE I.

Number of taxa	Recombination rate ρ	Mutation rate θ	recHMM correlation	SERES+recHMM correlation
5	0.5	0.5	-0.571	-0.692
5	0.5	1	-0.526	-0.676
5	0.5	2	-0.525	-0.651
5	1	0.5	-0.597	-0.675
5	1	1	-0.569	-0.678
5	1	2	-0.618	-0.684
5	2	0.5	-0.506	-0.661
5	2	1	-0.543	-0.665
5	2	2	-0.56	-0.648

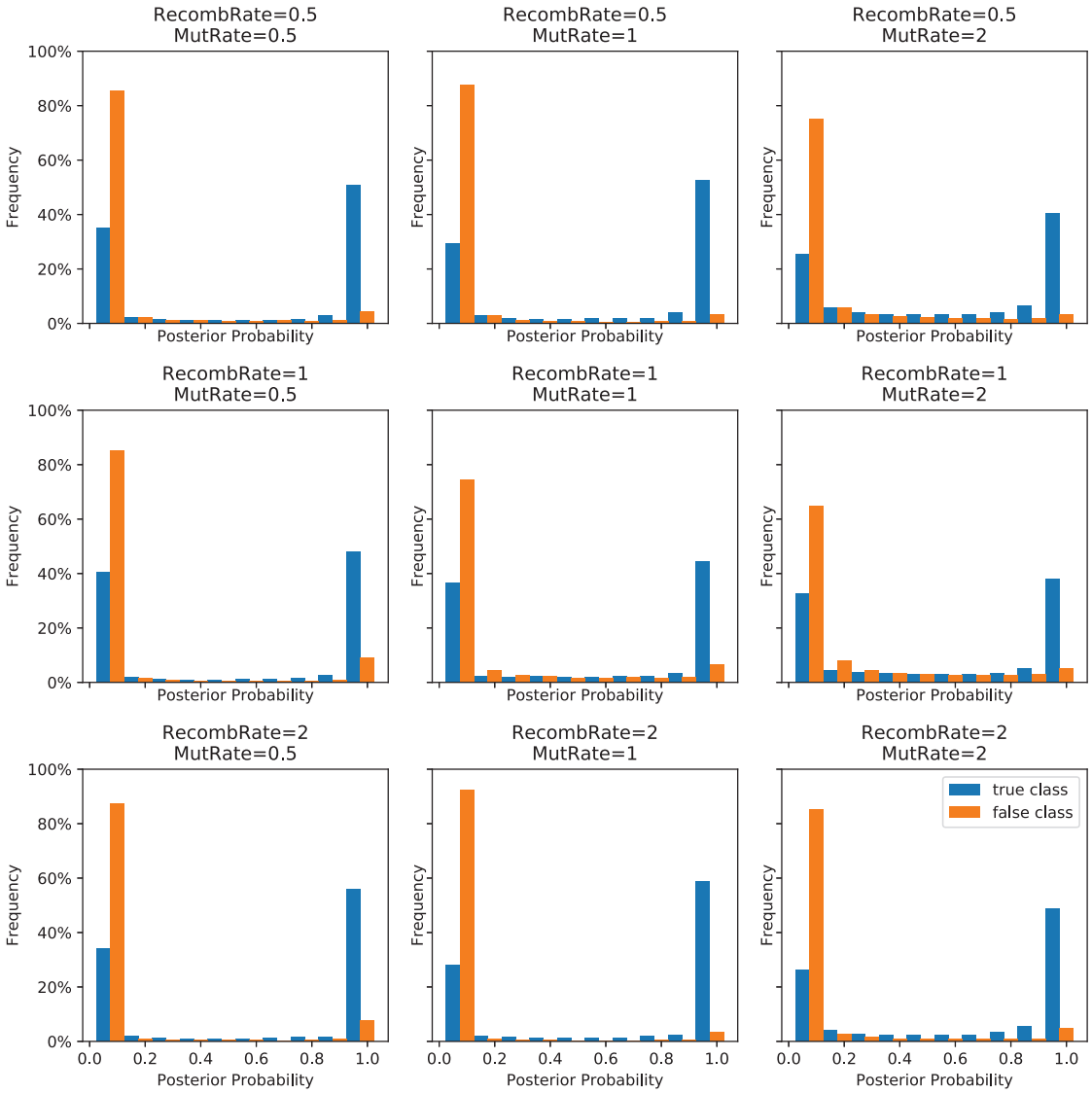

Fig. 1. **Histogram of posterior probabilities inferred by standalone recHMM method on 4-taxon model conditions.** Local gene tree topologies at a site were split into two classes: the "true class" consists of the true gene tree topology for the site, and the "false class" contains all other gene tree topologies. For each class and each replicate dataset in a model condition, the inferred posterior probabilities for gene trees at any site were binned into deciles; the resulting histogram was then normalized ($n = 30$). The normalized histograms for the true and false classes are shown in blue and orange, respectively.

learning optimization as the number of taxa increases. It is likely that conservatively limiting SERES-based re-estimation to 10 learning iterations is insufficient for the larger model conditions in our study. More intensive learning optimization may yield improved re-estimation and a greater performance benefit from augmenting recHMM with SERES.

The Appendix includes additional experiments that evaluate the impact of key method parameters. Supplementary Table S3 compares inference accuracy for recHMM versus SERES+recHMM as different choices are used for the SERES reversal probability γ. We found that the performance advantage returned by SERES+recHMM over standalone recHMM was robust to the choice of reversal probability γ so long as the chosen value was not too high; reasonable choices are equivalent to reversal breakpoints separated by at least 100 bp of sequence length on average. The results are consistent with the original motivation for sequence-aware resampling and re-estimation. [Wang et al., 2018] noted the correspondence between an rth order Markov process and a SERES random walk with reversal probability γ. For $\gamma = 0.5$, a first-order Markov process suffices; for $\gamma < 0.5$, higher-order Markovian processes are needed to capture sequential dependence. Essentially, smaller γ values mean that longer-distance sequential dependence is retained. Our results suggest that there is a critical point: past a certain threshold, longer-distance sequential dependence is critical to the performance of resampling and re-estimation for sequence-based inference problems. Supplementary Table S4 shows results for recHMM

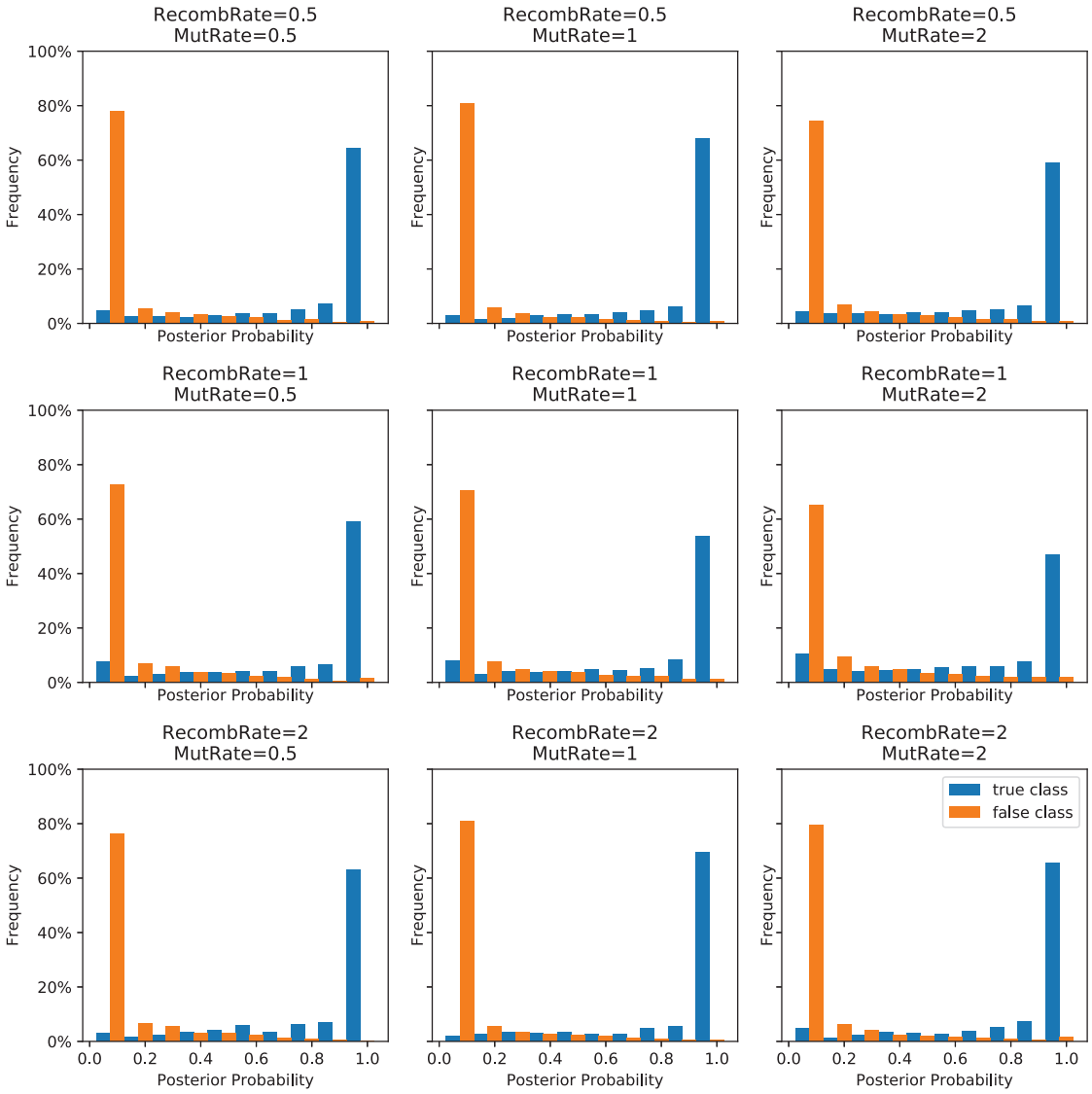

Fig. 2. **Histogram of posterior probabilities inferred by SERES+recHMM method on 4-taxon model conditions.** Figure layout and description are otherwise identical to Figure 1.

TABLE III
On 6-taxon model conditions, posterior probabilities inferred using SERES+recHMM were more highly correlated with topological accuracy compared to standalone recHMM. OTHERWISE, TABLE LAYOUT AND DESCRIPTION ARE IDENTICAL TO TABLE I.

Number of taxa	Recombination rate ρ	Mutation rate θ	recHMM correlation	SERES+recHMM correlation
6	0.5	0.5	-0.3312	-0.494
6	0.5	1	-0.251	-0.457
6	0.5	2	-0.360	-0.472
6	1	0.5	-0.376	-0.486
6	1	1	-0.469	-0.473
6	1	2	-0.507	-0.535
6	2	0.5	-0.382	-0.506
6	2	1	-0.504	-0.515
6	2	2	-0.455	-0.554

and SERES+recHMM analyses using alternative settings for the HMM state space size ϕ. We note that the simulated datasets in our study included between approximately 3 and 6 recombination-free intervals with distinct true gene trees, on average. Consistent with [Westesson and Holmes, 2009], we found that using a more complex recHMM model than necessary (i.e., more HMM states than the number of local gene trees encoded within a simulation replicate's ancestral recombination graph) resulted in overfitting. Interestingly, SERES+recHMM's performance was relatively robust to overfitting. A similar phenomenon is observed when using related data perturbation techniques to control overfitting (e.g., cross-validation). Finally, runtime and memory usage results are reported in Supplementary Table S2. Average runtime for the two methods were roughly comparable: average runtime

differences between the two methods were less than an hour on the 4-taxon and 5-taxon model conditions and less than two hours on the 6-taxon model conditions, and neither method consistently returned faster average runtime. Throughout our study, we observed low memory usage for both methods that amounted to less than 100 MiB.

Empirical study

As in the earlier studies of Lole et al. [1999] and Westesson and Holmes [2009], our SERES-based re-analysis clearly detected local topology switching that is consistent with historical recombination. The finding supports the hypothesis that the sequence 95IN21301 is recombinant.

As shown in Figure 3, the SERES+recHMM method recovered the five breakpoints described by both Lole et al. [1999] and Westesson and Holmes [2009]; the specific coordinates described in the latter study were 6402 bp, 6969 bp, 7073 bp, 9431 bp, and 9585 bp. In our re-analysis, the breakpoints correspond to switching between the blue topology and orange topology. SERES+recHMM posterior decoding also clearly showed inference uncertainty in the first few hundred bp of the input alignment.

Furthermore, Westesson and Holmes [2009] reported two additional breakpoints at 4328 bp and 4401 bp that were not described in the study of Lole et al. [1999]. Neither the standalone recHMM method nor the SERES+recHMM method recovered local topological incongruence in this specific region, although standalone recHMM posterior decoding recovered nearby local topology switching in the region from 4000 bp to 4200 bp. However, the SERES-based method indicated more uncertainty regarding gene tree inference within this region, relative to the five breakpoints described by both of the previous studies.

Re-analysis using SERES+recHMM also clarified patterns of local topology switching in other genomic regions. We detected local topological incongruence within the region from 3000 bp to 3500 bp. Some genomic regions exhibited local topology switching in standalone recHMM posterior decoding analysis that were not supported by the SERES+recHMM analysis (e.g., the region from 6000 to 6500 bp). Finally, throughout much of the genome alignment, SERES+recHMM inferred lower poster probability for the gene tree topology shown in green, relative to standalone recHMM. The region located between 5000 and 8000 bp was particularly striking: within this region, SERES+recHMM inferred basically zero probability for the topology shown in green, whereas recHMM inferred highly variable probability that was often far from zero.

IV. Conclusions

This study introduced the first application of SERES random walks on aligned sequences. The application is also the first to utilize SERES as a data perturbation technique to improve statistical inference and learning. Our performance validation experiments showed that coupling SERES with recHMM, an HMM-based method for recombination-aware local genealogical inference, yielded improved local inferences and potentially reduced type I and/or type II error. A re-analysis of an HIV genome sequence dataset clarifies the findings in the earlier study of Westesson and Holmes [2009].

We conclude with thoughts on future research. First, we note that statistical learning was a major bottleneck for the methods under study, particular for the SERES-based pipeline since optimization-based learning must be addressed for all SERES replicates. This scalability challenge is well suited to "pleasantly" parallel computation as well as more sophisticated parallelization techniques. Second, other studies have investigated ancestral recombination inference problems other than local genealogical inference (e.g., recombination rate estimation [Stumpf and McVean, 2003], recombination hotspot/coldspot detection [Auton and McVean, 2007, Myers et al., 2005], etc.). SERES resampling and re-estimation may prove to be similarly beneficial in these other contexts. Finally, we believe that we have only begun to realize the full potential of SERES random walks. As with other non-parametric and semi-parametric resampling techniques, SERES promises to find wide utility in computational biology/bioinformatics and beyond.

Acknowledgment

The authors gratefully acknowledge the support of the U.S. National Science Foundation (grant nos. CCF-1565719, CCF-1714417, DEB-1737898, and IOS-1740874 to KJL) and Michigan State University (faculty startup funds to KJL). All computational experiments and analyses were performed on the High Performance Computing Center (HPCC) cluster at Michigan State University.

References

A. Auton and G. McVean. Recombination rate estimation in the presence of hotspots. *Genome Research*, 17(8):1219–1227, 2007.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2): 123–140, 1996.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, pages 54–75, 1986.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural em algorithm for phylogenetic inference. *Journal of Computational Biology*, 9(2):331–353, 2002.

J. Hein, M. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*. Oxford University Press, Oxford, 2004.

R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

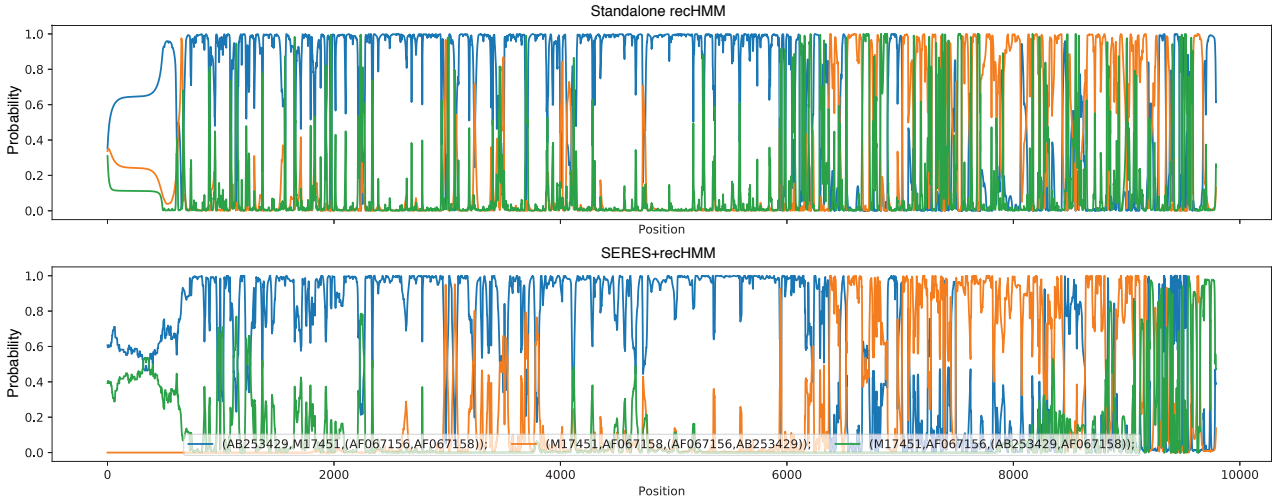

Fig. 3. **Posterior probabilities of local gene tree topologies inferred by standalone recHMM versus SERES+recHMM method on Indian HIV-1 dataset.** We re-analyzed a subset of the Indian HIV-1 genome dataset that was published by Lole et al. [1999]; Westesson and Holmes [2009] re-analyzed the original dataset using recHMM. Our re-analysis compared local gene tree probabilities computed using standalone recHMM posterior decoding (top panel) versus SERES+recHMM posterior decoding (bottom panel). The plots show posterior decoding probabilities (y-axis) versus genome coordinate (x-axis). Local gene tree probabilities are colored based on the three possible unrooted topologies for the four-taxon dataset (shown in either blue, orange, or green).

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2): 337–338, 2002.

D. Husmeier and F. Wright. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology*, 8(4):401–427, 2001.

T. Jukes and C. Cantor. *Evolution of Protein Molecules*, pages 21–132. Academic Press, New York, NY, USA, 1969.

G. Landan and D. Graur. Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution*, 24(6):1380–1383, 2007.

K. J. Liu, J. Dai, K. Truong, Y. Song, M. H. Kohn, and L. Nakhleh. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, 10(6):e1003649, 06 2014.

K. S. Lole, R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of Virology*, 73(1):152–160, 1999.

T. Mailund, J. Y. Dutheil, A. Hobolth, G. Lunter, and M. H. Schierup. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genetics*, 7(3):e1001319, 03 2011.

G. A. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

A. Rambaut and N. C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.

S. J. Russell and P. Norvig. *Artificial Intelligence: a Modern Approach*. Pearson Education Limited, 2016.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

M. P. Stumpf and G. A. McVean. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, 4(12):959, 2003.

W. Wang, J. Smith, H. A. Hejase, and K. J. Liu. Non-parametric and semi-parametric support estimation using sequential resampling random walks on biomolecular sequences. In M. Blanchette and A. Ouangraoua, editors, *Comparative Genomics*, pages 294–308. Springer International Publishing, 2018.

O. Westesson and I. Holmes. Accurate detection of recombinant breakpoints in whole-genome alignments. *PLoS Comput Biol*, 5(3):e1000318, 03 2009.

Additional Files

Additional file 1 — Appendix with Supplementary Material