1-2010

Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci

Elchanan Mossel
University of Pennsylvania

Sébastien Roch

Recommended Citation

Mossel, E., & Roch, S. (2010). Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7* (1), 166-171. http://dx.doi.org/10.1109/TCBB.2008.66

Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci

Abstract
We introduce a simple computationally efficient algorithm for reconstructing phylogenies from multiple gene trees in the presence of incomplete lineage sorting, that is, when the topology of the gene trees may differ from that of the species tree. We show that our technique is statistically consistent under standard stochastic assumptions, that is, it returns the correct tree given sufficiently many unlinked loci. We also show that it can tolerate moderate estimation errors.

Keywords
bioinformatics, estimation theory, genetics, stochastic processes, estimation errors, incomplete lineage sorting, multiple gene trees, multiple loci, phylogeny estimation, stochastic assumptions, biology and genetics, incomplete lineage sorting, probability and statistics, coalescent process, phylogenetics, population genetics, statistical consistency, topological concordance, algorithms, animals, chromosome mapping, computer simulation, evolution, humans, linkage, disequilibrium, models

Disciplines
Computer Sciences | Genetics and Genomics | Physical Sciences and Mathematics | Statistics and Probability

Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci*

Elchanan Mossel
Department of Statistics
University of California, Berkeley
mossel@stat.berkeley.edu

Sebastien Roch
Theory Group
Microsoft Research
Sebastien.Roch@microsoft.com

February 14, 2013

Abstract

We introduce a simple algorithm for reconstructing phylogenies from multiple gene trees in the presence of incomplete lineage sorting, that is, when the topology of the gene trees may differ from that of the species tree. We show that our technique is statistically consistent under standard stochastic assumptions, that is, it returns the correct tree given sufficiently many unlinked loci. We also show that it can tolerate moderate estimation errors.

1 Introduction

Phylogenies—the evolutionary relationships of a group of species—are typically inferred from estimated genealogical histories of one or several genes (or *gene trees*) [Fel04, SS03]. Yet it is well known that such gene trees may provide misleading information about the phylogeny (or *species tree*) containing them. Indeed, it was observed early on that a gene tree may be topologically inconsistent with its species tree, a phenomenon known as *incomplete lineage sorting*. See e.g. [Mad97, Nic01, Fel04] and references therein. Such discordance plays little role in the reconstruction of deep phylogenetic branchings but it is critical in the study of recently diverged populations [LP02, HM03, Kno04].

Two common approaches to deal with this issue are *concatenation* and *majority voting*. In the former, one concatenates the sequences originating from several

*Keywords: incomplete lineage sorting, gene tree, species tree, coalescent, topological concordance, statistical consistency. E.M. is supported by an Alfred Sloan fellowship in Mathematics and by NSF grants DMS-0528488, and DMS-0548249 (CAREER) and by ONR grant N0014-07-1-05-06.

genes and hopes that a tree inferred from the combined data will produce a better estimate. This approach appears to give poor results [KD07]. Alternatively, one can infer multiple gene trees and output the most common reconstruction (that is, take a majority vote). This is also often doomed to failure. Indeed, a recent, striking result of Degnan and Rosenberg [DR06] shows that, under appropriate conditions, the *most likely* gene tree may be inconsistent with the species tree; and this situation may arise on *any* topology with at least 5 species. See also [PN88, Tak89] for related results.

Other techniques are being explored that attempt to address incomplete lineage sorting, notably Bayesian [ELP07] and likelihood [SR07] methods. However the problem is still far from being solved as discussed in [MK06]. Here we propose a simple technique—which we call Global LAteSt Split or GLASS—for estimating species trees from multiple genes (or *loci*). Our technique develops some of the ideas of Takahata [Tak89] and Rosenberg [Ros02] who studied the properties of gene trees in terms of the corresponding species tree. In our main result, we show that GLASS is *statistically consistent*, that is, it always returns the correct topology given sufficiently many (unlinked) genes—thereby avoiding the pitfalls highlighted in [DR06]. We also obtain explicit convergence rates under a standard model based on Kingman's coalescent [Kin82]. Moreover, we allow the use of several alleles from each population and we show how our technique leads to an extension of Rosenberg's *topological concordance* [Ros02] to multiple loci.

We note the recent results of Steel and Rodrigo [SR07] who showed that Maximum Likelihood (ML) is statistically consistent under slightly different assumptions. An advantage of GLASS over likelihood (and Bayesian) methods is its computational efficiency, as no efficient algorithm for finding ML trees is known. Furthermore, GLASS gives explicit convergence rates—useful in assessing the quality of the reconstruction.

For more background on phylogenetic inference and coalescent theory, see e.g. [Fel04, SS03, HSW05, Nor01, Tav04].

Organization. The rest of the paper is organized as follows. We begin in Section 2 with a description of the basic setup. The GLASS method is introduced in Section 3. A proof of its consistency can be found in Sections 4 and 5. We show in Section 6 that GLASS remains consistent under moderate estimation errors. Finally in Section 7 we do away with the molecular clock assumption and we show how our technique can be used in conjunction with any distance matrix method.

2 Basic Setup

We introduce our basic modelling assumptions. See e.g. [DR06].

Species tree. Consider n isolated populations with a common evolutionary history given by the *species tree* $S = (V, E)$ with leaf set L. Note that $|L| = n$. For each branch e of S, we denote:

- N_e, the (haploid) population size on e (we assume that the population size remains constant along the branch);
- t_e, the number of generations encountered on e;
- $\tau_e = \frac{t_e}{2N_e}$, the length of e in standard coalescent time units;
- $\mu = \min_e \tau_e$, the shortest branch length in S.

The model does not allow migration between contemporaneous populations. Often in the literature, the population sizes $\{N_e\}_{e\in E}$, are taken to be equal to a constant N. Our results are valid in a more general setting.

Gene trees. We consider k *loci* $\mathcal{I}$. For each population l and each locus i, we sample a set of alleles $\mathcal{M}_l^{(i)}$. Each locus $i \in \mathcal{I}$ has a genealogical history represented by a *gene tree* $\mathcal{G}^{(i)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)})$ with leaf set $\mathcal{L}^{(i)} = \cup_l \mathcal{M}_l^{(i)}$. For two leaves a, b in $\mathcal{G}^{(i)}$, we let $\mathcal{D}_{ab}^{(i)}$ be the time in number of generations to the most recent common ancestor of a and b in $\mathcal{G}^{(i)}$. Following [Tak89, Ros02] we are actually interested in *interspecific* coalescence times. Hence, we define, for all $r, s \in L$,

$$\mathcal{D}_{rs}^{(i)} = \min\left\{\mathcal{D}_{ab}^{(i)} \ : \ a \in \mathcal{M}_r^{(i)}, b \in \mathcal{M}_s^{(i)}\right\}.$$

Inference problem. We seek to solve the following inference problem. We are given k gene trees as above, including accurate estimates of the coalescence times

$$\left\{\left(\mathcal{D}_{ab}^{(i)}\right)_{a,b\in\mathcal{L}^{(i)}}\right\}_{i\in\mathcal{I}}.$$

Our goal is to infer the species tree S.

Stochastic Model. In Section 4, we will first state the correcteness of our inference algorithm in terms of a combinatorial property of the gene trees. In Section 5, we will then show that under the following standard stochastic assumptions, this property holds for a moderate number of genes.

Namely, we will assume that each gene tree $\mathcal{G}^{(i)}$ is distributed according to a standard *coalescent process*: looking backwards in time, in each branch any two alleles coalesce at exponential rate 1 independently of all other pairs; whenever two populations merge in the species tree, we also merge the allele sets of the corresponding populations (that is, the coalescence proceeds on the *union* of both allele sets). We further assume that the k loci $\mathcal{I}$ are *unlinked* or in other words that the gene trees $\{\mathcal{G}^{(i)}\}_{i\in\mathcal{I}}$ are mutually independent.

Under these assumptions, an inference algorithm is said to be *statistically consistent* if the probability of returning an incorrect reconstruction goes to 0 as k tends to $+\infty$.

3 Species Tree Estimation

We introduce a technique which we call the Global LAteSt Split (GLASS) method.

Inference method. Consider first the case of a single gene ($k = 1$). Looking backwards in time, the first speciation occurs at some time T_1, say between populations r_1 and s_1. It is well known that, for any sample a from $\mathcal{M}_{r_1}^{(1)}$ and b from $\mathcal{M}_{s_1}^{(1)}$, the coalescence time $\mathcal{D}_{ab}^{(1)}$ between alleles a and b *overestimates* the divergence time of the populations. As noted in [Tak89], a better estimate of T_1 can be obtained by taking the smallest interspecific coalescence time between alleles in $\mathcal{M}_{r_1}^{(1)}$ and in $\mathcal{M}_{s_1}^{(1)}$, that is, by considering instead $\mathcal{D}_{r_1 s_1}^{(1)}$.

The inference then proceeds as follows. First, cluster the two populations, say r_1 and s_1, with smallest interspecific coalescence time $\mathcal{D}_{r_1 s_1}^{(1)}$. Define the coalescence time of two clusters $A, B \subseteq L$ as the minimum interspecific coalescence time between populations in A and in B, that is,

$$\mathcal{D}_{AB}^{(1)} = \min\left\{\mathcal{D}_{rs}^{(1)} \ : \ r \in A, s \in B\right\}.$$

Then, repeat as above until there is only one cluster left. This is essentially the algorithm proposed by Rosenberg [Ros02]. In particular, Rosenberg calls the implied topology on the populations so obtained the *collapsed gene tree*.

How to extend this algorithm to $k > 1$? As we discussed earlier, one could infer a gene tree as above for each locus and take a majority vote—but this approach fails [DR06]; in particular, it is generally not statistically consistent.

Another natural idea is to get a "better" estimate of coalescence times by *averaging* across loci. This leads to the Shallowest Divergence Clustering method of Maddison and Knowles [MK06]. We argue that a better choice is, instead, to take the *minimum* across loci. In other words, we apply the clustering algorithm above to the quantity

$$\mathcal{D}_{AB} = \min\left\{\mathcal{D}_{AB}^{(i)} \ : \ i \in \mathcal{I}\right\},$$

for all $A, B \subseteq L$ with $A \cap B = \emptyset$. The reason we consider the minimum is similar to the case of one locus and several samples per population above: it suffices to have *one* pair $a \in \mathcal{M}_r^{(i)}$, $b \in \mathcal{M}_s^{(i)}$ (for some i) with coalescence time T across *all pairs of samples in populations r and s (one from each)* and *all loci in $\mathcal{I}$* to provide indisputable evidence that the corresponding species branch before time T (looking backwards in time). In a sense, we build the "minimal" tree on L that is "consistent" with the evidence provided by the gene trees. This type of approach is briefly discussed by Takahata [Tak89] in the simple case of three populations (where the issues raised by [DR06] do not arise).

The algorithm, which we name GLASS, is detailed in Figure 1. We call the tree so obtained the *glass tree*. We show in the next section that GLASS is in fact statistically consistent.

Algorithm GLASS
Input: Gene trees $\{\mathcal{G}^{(i)}\}_{i\in\mathcal{I}}$ and coalescence times $\mathcal{D}_{ab}^{(i)}$ for all $i \in \mathcal{I}$ and $a, b \in \mathcal{L}^{(i)}$;
Output: Estimated topology S';

- [Intercluster coalescences] For all $A, B \subseteq L$ with $A \cap B = \emptyset$, compute

$$\mathcal{D}_{AB} = \min\left\{\mathcal{D}_{ab}^{(i)} \ : \ i \in \mathcal{I}, r \in A, s \in B, a \in \mathcal{M}_r^{(i)}, b \in \mathcal{M}_s^{(i)}\right\};$$

- [Clustering] Set $Q := \{\{r\} \ : \ r \in L\}$; Until $|Q| = 1$:
  - Denote the current partition $Q = \{A_1, \ldots, A_z\}$;
  - Let A', A'' minimize $\mathcal{D}_{AB}$ over all pairs $A, B \in Q$ (break ties arbitrarily);
  - Merge A' and A'' in Q;
- [Output] Return the topology implied by the steps above.

Figure 1: Algorithm GLASS.

Multilocus concordance. A gene tree with one sample per population is said to be *concordant* (sometimes also "congruent" or "consistent") with a species tree if

their (leaf-labelled) topologies agree. When the number of samples per population is larger than one, one cannot directly compare the topology of the gene tree with that of the species tree since they contain a different number of leaves. Instead, Rosenberg [Ros02] defines a gene tree to be *topologically concordant* with a species tree if the *collapsed gene tree* (see above) coincides with the species tree.

We extend Rosenberg's definition to multiple loci. We say that a collection of gene trees $\{\mathcal{G}^{(i)}\}_{i\in\mathcal{I}}$ is *multilocus concordant* with a species tree S if the *glass tree* agrees with the species tree. Therefore, to prove that GLASS is statistically consistent, it suffices to show that the probability of multilocus concordance goes to 1 as the number of loci goes to $+\infty$.

4 Sufficient Conditions

In this section, we state a simple combinatorial condition guaranteeing that GLASS returns the correct species tree. Our condition is an extension of Takahata's condition in the case of a single gene [Tak89]. See also [Ros02].

As before, let S be a species tree and $\{\mathcal{G}^{(i)}\}_{i\in I}$ a collection of gene trees. For a subset of leaves $A \subseteq L$, denote by $\langle A\rangle$ the most recent common ancestor (MRCA) of A in S. For a (internal or leaf) node v in S, we use the following notation:

- $\lfloor v \rfloor$ are the descendants of v in L;
- $\underline{t}_v$ is the time elapsed in number of generations between v and $\lfloor v \rfloor$;
- $\overline{t}_v$ is the time elapsed in number of generations between the immediate ancestor of v and $\lfloor v \rfloor$.

In particular, note that if e is the branch immediately above v, then we have

$$t_e = \overline{t}_v - \underline{t}_v.$$

Also, we call the subtree below v, *clade* v.

Our combinatorial condition can be stated as follows:

$$(\star) \quad \forall u, v \in V,\ \underline{t}_{\langle \lfloor u\rfloor \cup \lfloor v\rfloor\rangle} \leq \mathcal{D}_{\lfloor u\rfloor \lfloor v\rfloor} < \overline{t}_{\langle \lfloor u\rfloor \cup \lfloor v\rfloor\rangle}.$$

In words, for any two clades u, v, there is at least one locus i and one pair of alleles a, b with a from clade u and b from clade v such that the lineages of a and b coalesce before the end of the branch above the MRCA of u and v. (The first inequality is clear by construction.) By the next proposition, condition $(\star)$ is sufficient for multilocus concordance. Note, however, that it is not necessary. Nevertheless note that, by design, GLASS always returns a tree, even when the condition is not satisfied.

Proposition 1 (Sufficient Condition) *Assume that* $(\star)$ *is satisfied. Then, GLASS returns the correct species tree. In other words, the gene trees* $\{\mathcal{G}^{(i)}\}_{i\in\mathcal{I}}$ *are multilocus concordant with the species tree* S.

Proof: Let Q be one of the partitions obtained by GLASS along its execution and let B be the newly created set in Q. We claim that, under $(\star)$, it must be the case that

$$B = \lfloor \langle B \rangle \rfloor. \tag{1}$$

That is, B is the set of leaves of a clade in the species tree S. The proposition follows immediately from this claim.

We prove the claim by induction on the execution time of the algorithm. Property (1) is trivially true initially. Assume the claim holds up to time T and let Q, as above, be the partition at time $T+1$. Note that B is obtained by merging two sets B' and B'' forming a partition of B. By induction, B' and B'' satisfy (1). Now, suppose by contradiction that B does not satisfy (1). Let $\langle B \rangle_{\swarrow}$ and $\langle B \rangle_{\searrow}$ be the clades immediately below $\langle B \rangle$ with corresponding leaf sets $C' = \lfloor \langle B \rangle_{\swarrow} \rfloor$ and $C'' = \lfloor \langle B \rangle_{\searrow} \rfloor$. By our induction hypothesis, each of B' and B'' must be contained in one of C' or C''. Say $B' \subseteq C'$ and $B'' \subseteq C''$ without loss of generality. Moreover, since B does not satisfy (1), one of the inclusions is strict, say $B' \subset C'$. But by $(\star)$, any set X in Q containing an element of $C' - B'$ has

$$\mathcal{D}_{B'X} < \overline{t}_{\langle B' \cup X \rangle} \leq \overline{t}_{\langle B \rangle_{\swarrow}} = \underline{t}_{\langle B \rangle} = \underline{t}_{\langle B' \cup B'' \rangle} \leq \mathcal{D}_{B'B''}. \tag{2}$$

To justify the first two inequalities above, note that X is contained in the partition at time T and therefore satisfies (1). In particular, by construction

$$B' \cup X \subseteq C'.$$

Hence by (2), GLASS would not have merged B' and B'', a contradiction. ■

5 Statistical Consistency

In this section, we prove the consistency of GLASS.

Consistency. We prove the following consistency result. Note that the theorem holds for any species tree—including the "anomaly zone" of Degnan and Rosenberg [DR06].

Proposition 2 (Consistency) *GLASS is statistically consistent.*

Proof: Throughout the proof, time runs *backwards* as is conventional in coalescent theory. We use Proposition 1 and give a lower bound on the probability that condition $(\star)$ is satisfied.

Consider first the case of one locus and one sample per population. By $(\star)$, the reconstruction is correct if every time two populations meet, the corresponding alleles coalesce before the end of the branch immediately above. By classical coalescent calculations (e.g. [Tav84]), this happens with probability at least

$$(1 - e^{-\mu})^{n-1},$$

where we used the fact that there are $n - 1$ divergences.

Now consider the general case. Imagine running the coalescent processes of all loci *simultaneously*. Consider any branching between two populations. In every gene tree separately, if several alleles emerge on either sides of the branching, choose arbitrarily one allele from each side. The probability that the chosen allele pairs fail to coalesce before the end of the branch above in *all* loci is at most $e^{-k\mu}$ by independence. Indeed, irrespective of everything else going on, two alleles meet at exponential rate 1 (conditionally on the past). This finally gives a probability of success of at least

$$(1 - e^{-k\mu})^{n-1}.$$

For n and μ fixed, we get

$$(1 - e^{-k\mu})^{n-1} \to 1,$$

as $k \to +\infty$, as desired. ■

Rates. Implicit in the proof of Proposition 2 is the following convergence rate.

Proposition 3 (Rate) *It holds that*

$$\mathbb{P}[\text{ Multilocus Discordance }] \leq (n-1)[e^{-\mu}]^k.$$

In particular, for any $\varepsilon > 0$, taking

$$k = \frac{1}{\mu} \ln \left(\frac{n-1}{\varepsilon} \right),$$

we get

$$\mathbb{P}[\text{ Multilocus Discordance }] \leq \varepsilon.$$

Proof: Note that

$$1 - (1 - e^{-k\mu})^{n-1} \leq (n-1)[e^{-\mu}]^k.$$

■

Multiple alleles v. multiple loci. It is interesting to compare the relative effects of adding more alleles or more loci on the accuracy of the reconstruction. The result in Proposition 3 does not address this question. In fact, it is hard to obtain useful analytic expressions for small numbers of genes and alleles. However, the asymptotic behavior is quite clear. Indeed, as was pointed out in [Ros02] (see also [MK06] for empirical evidence), the benefit of adding more alleles eventually wears out. This is because the probability of observing any given number of alleles at the top of a branch is uniformly bounded in the number alleles existing at the bottom. More precisely, we have the following result which is to be contrasted with Proposition 3.

Proposition 4 (Multiple Alleles: Saturation Effect) *Let S be any species tree on n populations. Then, there is a $0 < q^* < 1$ (depending only on S) such that for any number of loci $k > 0$ and any number of alleles sampled per population, we have*

$$\mathbb{P}[\text{ Multilocus Discordance }] \geq (q^*)^k > 0.$$

In particular, for a fixed number of loci $k > 0$, as the number of alleles per populaiton goes to $+\infty$, the probability that GLASS correctly reconstructs S remains bounded away from 1.

Proof: Take any three populations a, b, c from S. Assume that a and b meet T_1 generations back and that c joins them T_2 generations later. For $w = a, b, c$ and $i \in \mathcal{I}$, let $Y_w^{(i)}$ be the event that in locus i there is only one allele remaining at the top of the branch immediately above w. Let $Z^{(i)}$ be the event that the topology of gene tree i restricted to $\{a, b, c\}$ is topologically discordant with S. It follows from bound (6.5) in [Tav84] that there is $0 < q' < 1$ independent of h such that

$$\mathbb{P}[Y_w^{(i)}] \geq q',$$

for all $i \in \mathcal{I}$ and $w \in \{a, b, c\}$. Also, it is clear that there is $0 < q'' < 1$ depending on T_2 such that

$$\mathbb{P}[Z^{(i)} \,|\, Y_w^{(i)},\ \forall w \in \{a, b, c\}] \geq q'',$$

for all $i \in \mathcal{I}$. Therefore, by independence of the loci, we have

$$\begin{aligned}
&\mathbb{P}[\text{ Multilocus Discordance }] \\
&\qquad \geq \prod_{i \in \mathcal{I}} \mathbb{P}[Z^{(i)} \,|\, Y_w^{(i)},\ \forall w \in \{a, b, c\}] \prod_{w \in \{a,b,c\}} \mathbb{P}[Y_w^{(i)}] \\
&\qquad \geq ((q')^3 q'')^k.
\end{aligned}$$

Take $q^* = (q')^3 q''$. That concludes the proof. ■

6 Tolerance to Estimation Error

The results of the previous section are somewhat unrealistic in that they assume that GLASS is given *exact* estimates of coalescence times. In this section, we relax this assumption.

Assume that the input to the algorithm is now a set of *estimated* coalescence times

$$\left\{ \left(\widehat{\mathcal{D}}_{ab}^{(i)} \right)_{a,b \in \mathcal{L}^{(i)}} \right\}_{i \in \mathcal{I}},$$

and, for all $A, B \subseteq L$, let

$$\widehat{\mathcal{D}}_{AB} = \min \left\{ \widehat{\mathcal{D}}_{ab}^{(i)} \ : \ i \in \mathcal{I}, r \in A, s \in B, a \in \mathcal{M}_r^{(i)}, b \in \mathcal{M}_s^{(i)} \right\},$$

be the corresponding estimated intercluster coalescence times computed by GLASS. Assume further that there is a $\delta > 0$ such that

$$\left| \widehat{\mathcal{D}}_{ab}^{(i)} - \mathcal{D}_{ab}^{(i)} \right| \leq \delta,$$

for all $i \in \mathcal{I}$ and $a, b \in \mathcal{L}^{(i)}$. In particular, note that

$$\left| \widehat{\mathcal{D}}_{AB} - \mathcal{D}_{AB} \right| \leq \delta,$$

for all $A, B \subseteq L$.

Let m be the shortest branch length in number of generations, that is,

$$m = \min\{t_e \ : \ e \in E\}.$$

We extend our combinatorial condition $(\star)$ to

$$(\hat{\star}) \quad \forall u, v \in V, \ \underline{t}_{\langle \lfloor u \rfloor \cup \lfloor v \rfloor \rangle} \leq \mathcal{D}_{\lfloor u \rfloor \lfloor v \rfloor} < \overline{t}_{\langle \lfloor u \rfloor \cup \lfloor v \rfloor \rangle} - 2\delta.$$

Then, we get the following.

Proposition 5 (Sufficient Condition:Noisy Case) *Assume that*

$$\delta < \frac{m}{2}, \tag{3}$$

and that $(\hat{\star})$ is satisfied. Then, GLASS returns the correct species tree.

Proof: The proof follows immediately from the argument in Proposition 1 by noting that equation (2) becomes

$$\begin{aligned}\widehat{\mathcal{D}}_{B'X} &\leq \mathcal{D}_{B'X} + \delta \\ &< \overline{t}_{\langle B' \cup X \rangle} - \delta \\ &\leq \underline{t}_{\langle B' \cup B'' \rangle} - \delta \\ &\leq \mathcal{D}_{B'B''} - \delta \\ &\leq \widehat{\mathcal{D}}_{B'B''}.\end{aligned}$$

Condition (3) ensures that $(\hat{\star})$ is satisfiable. ■

Moreover, we have immediately:

Proposition 6 (Consistency & Rate: Noisy Case) *Assume that*

$$\delta < \frac{m}{2}.$$

Then GLASS is statistically consistent. Moreover, let

$$\Lambda = \frac{m - 2\delta}{m},$$

then it holds that

$$\mathbb{P}[\text{ Incorrect Reconstruction }] \leq (n-1)[e^{-\mu\Lambda}]^k.$$

In particular, for any $\varepsilon > 0$, taking

$$k = \frac{1}{\mu\Lambda} \ln\left(\frac{n-1}{\varepsilon}\right),$$

we get

$$\mathbb{P}[\text{ Incorrect Reconstruction }] \leq \varepsilon.$$

7 Generalization

The basic observation underlying our approach is that distances between populations may be estimated correctly using the minimum divergence time among all individuals and all genes.

Actually, this observation may be used in conjunction with any distance-based reconstruction algorithm. (See e.g. [Fel04, SS03] for background on distance matrix methods.) This can be done under very general assumptions as we discuss

next. First, we do away with the molecular clock assumption. Indeed, it turns out that $\mathcal{D}_{ab}^{(i)}$ need not be the divergence time between a and b for gene i. Instead, we take $\mathcal{D}_{ab}^{(i)}$ to be the molecular distance between a and b in gene i, that is, the time elapsed from the divergence point to a and b integrated against the rate of mutation. We require that the rate of mutation be the same for all genes and all individuals in the same branch of the species tree, but we allow rates to differ across branches. Below, all quantities of the type $\mathcal{D}, \widehat{\mathcal{D}}$ etc. are given in terms of this molecular distance.

For any two clusters $A, B \subseteq L$, we define

$$\widehat{\mathcal{D}}_{AB} = \min\left\{\widehat{\mathcal{D}}_{ab}^{(i)} \ : \ i \in \mathcal{I}, r \in A, s \in B, a \in \mathcal{M}_r^{(i)}, b \in \mathcal{M}_s^{(i)}\right\}, \qquad (4)$$

as before. Let

$$m' = \min\{t_e \rho_e \ : \ e \in E\},$$

where ρ_e is the rate of mutation on branch e. It is easy to generalize condition $(\hat{\star})$ so that we can use (4) to estimate all molecular distances between pairs of populations up to an additive error of, say, $m'/4$. Then using standard four-point methods, we can reconstruct the species tree correctly.

Note furthermore that by the results of [ESSW99], it suffices in fact to estimate distances between pairs of populations that are "sufficiently close." We can derive consistency conditions which guarantee the reconstruction of the correct species tree in that case as well.

References

[DR06] J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5), May 2006.

[ELP07] Scott V. Edwards, Liang Liu, and Dennis K. Pearl. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941, 2007.

[ESSW99] P. L. Erdös, M. A. Steel;, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 1). *Random Struct. Algor.*, 14(2):153–184, 1999.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer, New York, New York, 2004.

[HM03] J. Hey and C. A. Machado. The study of structured populations–new hope for a difficult and divided science. *Nat. Rev. Genet.*, 4(7):535–543, July 2003.

[HSW05] Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory*. Oxford University Press, USA, February 2005.

[KD07] L. S. Kubatko and J. H. Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24, February 2007.

[Kin82] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, (Special Vol. 19A):27–43, 1982. Essays in statistical science.

[Kno04] L. L. Knowles. The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, 17(1):1–10, 2004.

[LP02] Knowles L. L. and Maddison W. P. Statistical phylogeography. *Mol. Ecol.*, 11(12):2623–35, 2002.

[Mad97] Wayne P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

[MK06] Wayne Maddison and L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30, February 2006.

[Nic01] R Nichols. Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16(7):358–364, July 2001.

[Nor01] M. Nordborg. Coalescent theory. In D. J. Baldingand M. J. Bishop and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 179–212. John Wiley & Sons, Inc., Chichester, U.K., 2001.

[PN88] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol. Biol. Evol.*, 5(5):568–583, September 1988.

[Ros02] N. A. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.*, 61(2):225–247, March 2002.

[SR07] M. Steel and Allen Rodrigo. Maximum likelihood supertrees. Preprint, 2007.

[SS03] C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003.

[Tak89] N. Takahata. Gene genealogy in three related population: Consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

[Tav84] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26(2):119–164, October 1984.

[Tav04] Simon Tavaré. Ancestral inference in population genetics. In *Lectures on probability theory and statistics*, volume 1837 of *Lecture Notes in Math.*, pages 1–188. Springer, Berlin, 2004.