Syst. Biol. 65(2):328–333, 2016

DOI:10.1093/sysbio/syv089
Advance Access publication November 28, 2015

Maximum Likelihood Phylogenetic Inference is Consistent on Multiple Sequence Alignments, with or without Gaps

Jakub Truszkowski[1,2,*] and Nick Goldman[1]

[1] *European Molecular Biology Laboratory, European Bioinformatics Institute Wellcome Genome Campus, Hinxton, CB10 1SD, UK;*
[2] *Cancer Research UK Cambridge Institute, University of Cambridge Robinson Way, Cambridge CB2 0RE, UK*
[*] *Correspondence to be sent to: European Molecular Biology Laboratory, European Bioinformatics Institute Wellcome Genome Campus, Hinxton, CB10 1SD, UK; E-mail: jakubt@ebi.ac.uk.*

Received 5 August 2015; reviews returned 17 November 2015; accepted 19 November 2015
Associate Editor: Edward Susko

Abstract.—We prove that maximum likelihood phylogenetic inference is consistent on gapped multiple sequence alignments (MSAs) as long as substitution rates across each edge are greater than zero, under mild assumptions on the structure of the alignment. Under these assumptions, maximum likelihood will asymptotically recover the tree with edge lengths corresponding to the mean number of substitutions per site on each edge. This refutes Warnow's recent suggestion (Warnow 2012) that maximum likelihood phylogenetic inference might be statistically inconsistent when gaps are treated as missing data, even if the MSA is correct. We also derive a simple new proof of maximum likelihood consistency of ungapped alignments. [Kullback–Leibler divergence, maximum likelihood, missing data, multiple sequence alignment, phylogeny.]

Phylogenetic inference is a fundamental bioinformatics problem. The amount of sequence data available is increasing and researchers design methods that try to extract all information from large data sets. One desirable property of a phylogenetic method is statistical consistency: as the number of sites tends to infinity, the inferred tree should converge to the true tree that generated the data. It is now widely accepted that standard phylogenetic methods, such as distance methods or maximum likelihood, achieve statistical consistency under Markov models of evolution without insertions and deletions (see Felsenstein (2004) or RoyChoudhury (2014) for a discussion).

In practice, phylogenies are usually built from multiple sequence alignments (MSAs) that contain insertions and deletions (indels). Most phylogenetic methods treat indels as missing data and do not attempt to model the indel process. It has been widely assumed that statistical consistency of standard phylogenetic methods extends to this setting, though to our knowledge no proof of this has been published. Recently, Warnow (2012) called this assumption into question by presenting an example of an evolutionary process where methods treating indels as missing data will fail to achieve consistency. More specifically, Warnow pointed out that for alignments that contain indels but no substitutions, methods that neglect indels will give equal support to all tree topologies, whereas an explicit model of the indel process can achieve statistical consistency. Although such evolutionary regimes are rare in nature, Warnow claimed that this example might be representative of a larger class of models where consistency may not hold, even when the alignment is correct.

Here, we prove that the "no substitutions" scenario highlighted by Warnow is in fact the only case where standard phylogenetic inference methods fail to achieve statistical consistency on correct MSAs. Specifically, we prove that for any evolutionary process where each edge in the tree has a nonzero substitution rate, standard maximum likelihood inference is consistent on MSAs, under minimal assumptions on the indel process.

Preliminaries

Trees, Alignments, and Substitutions

A *phylogenetic tree* T on n leaves is a tree (not necessarily binary) whose leaves are uniquely labeled with elements of the taxon set $N = \{1,\ldots,n\}$. Each edge in T has an associated positive length. For any two leaves a,b, the distance $d_T(a,b)$ is defined as the sum of lengths of all edges on the path from a to b. For a taxon set $S \subseteq N$, $T_{|S}$ is the unique tree obtained by removing all vertices of T that do not lie on a path between some two leaves in S, suppressing all degree-two vertices in the remaining tree, and adjusting the edge lengths so that $d_{T_{|S}}(a,b) = d_T(a,b)$ for every choice of a and b in S.

A *MSA* $\mathbb{A}$ is an $n \times m$ matrix with entries from the set $\Sigma = \{A,C,G,T,-\}$, where n is the number of taxa and m is the number of sites in the alignment. Characters in the same column of $\mathbb{A}$ are descended from a common ancestral site. $\mathbb{A}_{ij} = -$ represents a "gap" in the alignment, indicating that taxon i has no sites that are homologous to any site in column j. This can happen as a result of a deletion on a branch leading to taxon i or an insertion on any branch that does not lie between i and the root of the tree.

For a taxon set $S \subseteq N$, a *site category* $C(S)$ is the set of columns in $\mathbb{A}$ whose nongap characters occur precisely in taxa from S. For sets $X \subseteq N$ and $Y \subseteq \{1,\ldots,m\}$, we denote by $\mathbb{A}[X,Y]$ the restriction of $\mathbb{A}$ to rows in X and columns in Y. In particular, $\mathbb{A}[S,C(S)]$ is the restriction of $\mathbb{A}$ to the rows in S and columns in $C(S)$, and therefore contains no gaps. By slight abuse of notation, we will write $\mathbb{A}[*,j]$ to refer to the characters in the j-th column of $\mathbb{A}$.

We need to make some minimal assumptions about the evolutionary process in order to prove our main result. We assume the sequences evolve along the edges of the tree under a continuous-time Markov chain where substitutions and indels are independent of each other, starting from an unobserved ancestral sequence at the root of the tree. We also assume that for each insertion, the inserted nucleotides are drawn independently from the equilibrium distribution of the substitution model. For simplicity of presentation, we assume the Jukes–Cantor model for nucleotide substitution (Jukes and Cantor 1969), with substitutions independent between sites, though our proof could be generalized to more complicated models (see Section 5).

The assumption of independence between indel and substitution processes appears to be crucial. In a recent paper, McTavish et al. (2015) showed that distance methods are inconsistent when some sites are invariant with respect to both substitutions and indels. Other authors have also reported biases in reconstructed phylogenies when patterns of missing data were correlated with the substitution process (Grievink et al. 2013; Roure et al. 2013). Similar problems have been discussed in statistics literature (Allison 2001).

Given a tree T with specified branch lengths and an evolutionary model, we can compute the probability $\Pr[\mathbb{A}[*,j]|T]$ of a pattern of nucleotides arising at the leaves of T, for example using Felsenstein's pruning algorithm (Felsenstein 2004). In standard phylogenetic analyses, gaps at site j are treated as missing data. This is equivalent to computing the likelihood $\Pr[\mathbb{A}[S,j]|T_{|S}]$ where S is the set of taxa that do not have gaps at site j.

For a sequence alignment $\mathbb{A}$, the likelihood of $\mathbb{A}$ given tree T is the product of per-site likelihoods for each site in the alignment. Taking the logarithm of the likelihood, we can write

$$\log\Pr[\mathbb{A}|T] = \sum_{j=1}^{m} \log\Pr[\mathbb{A}[*,j]|T].$$

It is often convenient to use the normalized per site log-likelihood $\mathcal{L}(\mathbb{A}|T) = \frac{1}{m}\log\Pr[\mathbb{A}|T]$.

The *maximum likelihood phylogeny* is the tree that maximizes the likelihood of the alignment:

$$ML(\mathbb{A}) = \arg\max_{T} \log\Pr[\mathbb{A}|T] = \arg\max_{T} \mathcal{L}(\mathbb{A}|T).$$

Finally, we define a metric on the space of all phylogenies on N as follows: let $D(T_1,T_2) = \max_{a,b\in N}|d_{T_1}(a,b) - d_{T_2}(a,b)|$. It can be easily verified that this is indeed a metric, as long as all edge lengths are positive. We note that this is not true in the case discussed by Warnow, where all edges have zero length as the substitution rate is zero across the tree. When zero-length edges are allowed, it is possible to find trees T_1 and T_2 with distinct topologies such that $D(T_1,T_2) = 0$, which violates the definition of a metric.

Consistency of ML has been stated in several papers (Yang 1994; Rogers 1997). A formal statement of consistency can be written as follows:

Theorem 1 *Let $\mathbb{A}^m$ be an ungapped m-column MSA generated from tree T^*. Then $D(ML(\mathbb{A}^m),T^*) \to 0$ with probability 1 as $m\to\infty$.*

Several proofs of Theorem 1 have been proposed, but some were later found to be incorrect. The proofs of Yang (1994) and Rogers (1997) show that for any tree $T \neq T^*$, we have $\mathcal{L}(\mathbb{A}^m|T) < \mathcal{L}(\mathbb{A}^m|T^*)$ with probability 1 as $m\to\infty$. Unfortunately, this is insufficient to prove consistency, since it does not preclude the possibility that the sequence $ML(\mathbb{A}^1), ML(\mathbb{A}^2), ML(\mathbb{A}^3), \ldots$ does not converge to T^*. This was noticed recently by RoyChoudhury (2014), who provides a more detailed discussion of these two proofs. Felsenstein (1973) argued that consistency follows from the result by Wald (1949), who proved statistical consistency for maximum likelihood estimators under very general conditions. However, Felsenstein's claim was disputed by Yang (1994) and Farris (1999), since Wald's proof relies on several technical assumptions which are not straightforward to verify for phylogenetic trees. These arguments were in turn countered by more recent works Swofford et al. (2001); RoyChoudhury (2014). Chang (1996) proved that ML is consistent for a more complicated model where each branch in the tree has its own substitution matrix. Although Chang's model includes the basic time-homogeneous model as a special case, the proof of identifiability for Markov models of evolution is somewhat complicated due to its generality. Chang's proof also requires the true tree to be binary, though the author does mention that the result could be extended to nonbinary trees. To our knowledge, no simple, correct proof has been published for the basic scenario where sequences evolve according to a fixed time-reversible model, constant across sites and edges.

We will use the following property to establish our result:

Lemma 1. *For any $\epsilon > 0$, $\Pr[\sup_{T:D(T,T^*)>\epsilon}\mathcal{L}(\mathbb{A}^m|T) \geq \mathcal{L}(\mathbb{A}^m|T^*)] \to 0$ as $m\to\infty$.*

This states that as the number of alignment columns tends to infinity, it is almost certain that an ungapped MSA has a higher likelihood on T^* than any other tree T that is "distinct enough" from T^*. We prove this lemma towards the end of the article.

The Indel Process

We assume that the process starts from an ancestral sequence of length m_{anc}. The expected length of the alignment is proportional to m_{anc} and we have $m\to\infty$ whenever $m_{anc}\to\infty$. For simplicity, we will write $m\to\infty$ from now on. We make the following two mild assumptions about the structure of the alignment:

Assumption 1. For each pair a,b of taxa, let K_{ab} be the number of columns j such that $\mathbb{A}^m_{aj} \neq -$ and $\mathbb{A}^m_{bj} \neq -$. Then, with probability 1,

$$\min_{a,b} K_{ab} \to \infty \text{ as } m \to \infty.$$

Assumption 2. For each $S \subseteq N$, either $|C(S)| = 0$ or $|C(S)| \to \infty$ as $m \to \infty$, with probability 1.

Assumption 1 states that every pair of taxa has infinitely many shared nongapped sites as the size of the MSA tends to infinity. Assumption 2 states that every possible site category either is not observed or is observed infinitely many times as the size of the MSA tends to infinity. These two assumptions hold for a large class of Markov models of sequence evolution, including for example the TKF model (Thorne et al. 1991), the "long indel" model (Miklós et al. 2004) and the recently introduced Poisson indel process (Bouchard-Côté and Jordan 2013). These models assume that sequences evolve according to a continuous-time Markov chain where the insertion and deletion rates are uniform across the sequence. This uniformity directly implies Assumption 2. Assumption 1 is a consequence of the fact that, under these models, each site of the ancestral sequence has a positive probability of not undergoing a deletion between any two leaves of the tree, as long as indel rates on every branch are finite. Note that the only case when we have $|C(S)| = 0$ as $m \to \infty$ for some sets S occurs when there are no deletions. In that case, an insertion remains present in all descendant lineages. Thus, the sets for which $C(S) > 0$ coincide with the clades in the tree.

Consistency for Gapped Alignments

We can now state and prove consistency of ML inference of phylogeny for gapped alignments. The key intuition behind the proof is that the likelihood of a gapped alignment can be represented as a sum of likelihoods of a finite number of gapped alignments created by restricting $\mathbb{A}$ to different site categories. We use Lemma 1 to prove that consistency for these ungapped alignments is enough to ensure consistency of the whole gapped alignment.

Theorem 2 *Under Assumptions 1 and 2, maximum likelihood phylogenetic inference is consistent. More precisely, if $\mathbb{A}^m$ is an m-column alignment containing gaps, $D(ML(\mathbb{A}^m), T^*) \to 0$ with probability 1 as $m \to \infty$.*

Proof. The log-likelihood of the alignment given tree T can be written as the product of per-site likelihoods:

$$\begin{aligned}\mathcal{L}(\mathbb{A}^m|T) &= \frac{1}{m}\sum_{j=1}^{m} \log \Pr[\mathbb{A}^m[*,j]|T] \\ &= \frac{1}{m}\sum_{S \subseteq N} \log \Pr[\mathbb{A}^m[*,C(S)]|T].\end{aligned}$$

Since gaps are treated as missing data, we can write equivalently

$$\mathcal{L}(\mathbb{A}^m|T) = \frac{1}{m}\sum_{S \subseteq N} \log \Pr[\mathbb{A}^m[S,C(S)]|T_{|S}].$$

For each taxon set S, $\mathbb{A}^m[S,C(S)]$ is an ungapped alignment.

By Lemma 1, for all $\epsilon > 0$ we have

$$\Pr\left[\sup_{T:D(T_{|S},T^*_{|S})>\epsilon} \mathcal{L}(\mathbb{A}^m[S,C(S)]|T_{|S}) \geq \mathcal{L}(\mathbb{A}^m[S,C(S)]|T^*_{|S})\right] \to 0 \text{ as } m \to \infty \tag{1}$$

for all S such that $|C(S)| > 0$. By Assumption 1, for any tree T such that $D(T,T^*) > \epsilon$, there exists at least one set S such that $D(T_{|S}, T^*_{|S}) > \epsilon$ and $|C(S)| \to \infty$ as $m \to \infty$. We can use this to bound the probability of any tree distinct from T^* having higher likelihood:

$$\Pr\left[\sup_{T:D(T,T^*)>\epsilon} \mathcal{L}(\mathbb{A}^m|T) \geq \mathcal{L}(\mathbb{A}^m|T^*)\right] \leq \Pr\left[\bigcup_{S \subseteq N} \sup_{T:D(T_{|S},T^*_{|S})>\epsilon} \mathcal{L}(\mathbb{A}^m[S,C(S)]|T_{|S}) \geq \mathcal{L}(\mathbb{A}^m[S,C(S)]|T^*_{|S})\right].$$

By the union bound, iterating over all subsets S of N, we get

$$\Pr\left[\sup_{T:D(T,T^*)>\epsilon} \mathcal{L}(\mathbb{A}^m|T) \geq \mathcal{L}(\mathbb{A}^m|T^*)\right] \leq \sum_{S \subseteq N} \Pr\left[\sup_{T:D(T_{|S},T^*_{|S})>\epsilon} \mathcal{L}(\mathbb{A}^m[S,C(S)]|T_{|S}) \geq \mathcal{L}(\mathbb{A}^m[S,C(S)]|T^*_{|S})\right].$$

From Equation (1), we see that the above sum tends to zero as $m \to \infty$. It follows that $\Pr[D(ML(\mathbb{A}^m), T^*) \leq \epsilon] \to 1$ for all $\epsilon > 0$. Consequently, we have $D(ML(\mathbb{A}^m), T^*) \to 0$ as $m \to \infty$, as desired. ■

A New Proof of ML Consistency for Ungapped Alignments

In this section, we present a new and simple proof of Theorem 1, consistency for maximum likelihood on ungapped alignments.

A pattern at site i is a tuple $c = (c_1, \ldots, c_n) \in \Sigma^n$ denoting the characters at alignment positions $\mathbb{A}_{1j}, \mathbb{A}_{2j}, \ldots, \mathbb{A}_{nj}$, respectively. The probability f_c^* of seeing each c at a site is specified by tree T^*, and let $f_c = m_c/m$ be the empirical frequency of pattern c in the alignment. It is easy to see that $f_c \to f_c^*$ with probability 1 as $m \to \infty$. For a fixed pair of taxa a,b, we will write f_{xy}^{ab} for the fraction of columns in $\mathbb{A}$ such that $\mathbb{A}_{aj}=x$ and $\mathbb{A}_{bj}=y$, and denote by xyc' a pattern c such that $\mathbb{A}_{aj}=x$, $\mathbb{A}_{bj}=y$ and $c' \in \Sigma^{n-2}$ is a tuple containing the characters of c at taxa other than a and b. Likewise, we will write p_{xy}^d for the probability of jointly observing character x at a and character y at b given that $d_T(a,b)=d$.

The following lemma provides an upper bound on the log-likelihood of the data given a misspecified tree T.

Lemma 2. *Let T be a tree such that there exist $a,b \in N$ with $d_T(a,b)=d$. Then*

$$\mathcal{L}(\mathbb{A}|T) \leq \sum_{c\in\Sigma^n} f_c \log f_c - D_{KL}(\left\{f_{xy}^{ab}\right\} || \left\{p_{xy}^d\right\})$$

where $D_{KL}(\{f_{xy}^{ab}\}||\{p_{xy}^d\}) = \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log \frac{f_{xy}^{ab}}{p_{xy}^d}$ *is the Kullback-Leibler (KL) divergence between pairwise character distributions* $\{f_{xy}^{ab}\}$ *and* $\{p_{xy}^d\}$*,*

Proof. Let p_c be the probability of c under tree T. The normalized log-likelihood $\mathcal{L}(\mathbb{A}|T)$ can be written as

$$\mathcal{L}(\mathbb{A}|T) = \sum_{c\in\Sigma^n} f_c \log p_c.$$

Our goal is to bound $\mathcal{L}(\mathbb{A}|T)$ when T is constrained so that $d_T(a,b)=d$. We can write

$$\begin{aligned}\mathcal{L}(\mathbb{A}|T) &= \sum_{xy\in\Sigma^2}\sum_{c'\in\Sigma^{n-2}} f_{xy}^{ab} f'_{xyc'} \log p_{xy}^d p'_{xyc'} \\ &= \sum_{xy\in\Sigma^2} f_{xy}^{ab} \sum_{c'\in\Sigma^{n-2}} f'_{xyc'} (\log p_{xy}^d + \log p'_{xyc'}),\end{aligned}$$

where $f'_{xyc'} = f_{xyc'}/f_{xy}^{ab}$ and $p'_{xyc'} = p_{xyc'}/p_{xy}^d$. Since $\sum_{c'\in\Sigma^{n-2}} f_{xyc'} = f_{xy}^{ab}$, we have $\sum_{c'\in\Sigma^{n-2}} f'_{xyc'} = 1$ and hence

$$\mathcal{L}(\mathbb{A}|T) = \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^d + \sum_{xy\in\Sigma^2} f_{xy}^{ab} \sum_{c'\in\Sigma^{n-2}} f'_{xyc'} \log p'_{xyc'}. \tag{2}$$

To get an upper bound on $\mathcal{L}(\mathbb{A}|T)$, we find the values of $p'_{xyc'}$ that maximize the expression in Equation (2) subject to $\sum_{c'\in\Sigma^{n\text{âˆ’}2}} p'_{xyc'} = 1$ for each $xy \in \Sigma^2$. By Gibbs' inequality (MacKay 2003), this occurs when $p'_{xyc'} = f'_{xyc'}$ for all $xyc' \in \Sigma^n$, and therefore

$$\mathcal{L}(\mathbb{A}|T) \leq \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^d + \sum_{xy\in\Sigma^2} f_{xy}^{ab} \sum_{c'\in\Sigma^{n-2}} f'_{xyc'} \log f'_{xyc'}.$$

Using $\log f'_{xyc'} = \log f_{xyc'} - \log f_{xy}^{ab}$, we get

$$\begin{aligned}\mathcal{L}(\mathbb{A}|T) \leq{} & \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^d \\ & + \sum_{xy\in\Sigma^2} f_{xy}^{ab} \sum_{c'\in\Sigma^{n-2}} f'_{xyc'} (\log f_{xyc'} - \log f_{xy}^{ab}) \\ \implies \mathcal{L}(\mathbb{A}|T) \leq{} & \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^d + \sum_{xy\in\Sigma^2}\sum_{c'\in\Sigma^{n-2}} f_{xy}^{ab} f'_{xyc'} \log f_{xyc'} \\ & - \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log f_{xy}^{ab} \\ \implies \mathcal{L}(\mathbb{A}|T) \leq{} & \sum_{xy\in\Sigma^2} f_{xy}^{ab} (\log p_{xy}^d - \log f_{xy}^{ab}) + \sum_{c\in\Sigma^n} f_c \log f_c \\ \implies \mathcal{L}(\mathbb{A}|T) \leq{} & \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log \frac{p_{xy}^d}{f_{xy}^{ab}} + \sum_{c\in\Sigma^n} f_c \log f_c \\ ={} & \sum_{c\in\Sigma^n} f_c \log f_c - D_{KL}\left(\left\{f_{xy}^{ab}\right\} \middle|\middle| \left\{p_{xy}^d\right\}\right)\end{aligned}$$

as desired. ■

We note that the bound in Lemma 2 is quite loose, as it does not incorporate the constraint that the pattern probabilities p_c must be consistent with some phylogenetic tree T.

The next lemma ensures that, given enough data, there exist no high-likelihood trees that are bounded away from the true tree T^*:

Lemma 3. *Given any taxon pair a,b and any $\epsilon > 0$, and writing $d_{T^*}(a,b) = d$, we have*

$$\lim_{m\to\infty} \sup_{T:d_T(a,b)\geq d+\epsilon} \mathcal{L}(\mathbb{A}^m|T) \leq \sum_{c\in\Sigma^n} f_c^* \log f_c^* - D_{KL}\left(\left\{f_{xy}^{ab}\right\} \middle|\middle| \left\{p_{xy}^d\right\}\right) \tag{3}$$

and

$$\lim_{m\to\infty} \sup_{T:d_T(a,b)\leq d-\epsilon} \mathcal{L}(\mathbb{A}^m|T) \leq \sum_{c\in\Sigma^n} f_c^* \log f_c^* - D_{KL}\left(\left\{f_{xy}^{ab}\right\} \middle|\middle| \left\{p_{xy}^d\right\}\right) \tag{4}$$

with probability 1.

Proof. Applying Lemma 2 to all trees T such that $d_T(a,b) \geq d+\epsilon$, we get

$$\sup_{T:d_T(a,b)\geq d+\epsilon} \mathcal{L}(\mathbb{A}^m|T) \leq \sum_{c\in\Sigma^n} f_c \log f_c - \inf_{d'\geq d+\epsilon} D_{KL}\left(\left\{f_{xy}^{ab}\right\} \middle|\middle| \left\{p_{xy}^d\right\}\right)$$

which is equivalent to

$$\sup_{T:d_T(a,b)\geq d+\epsilon} \mathcal{L}(\mathbb{A}^m|T) \leq \sum_{c\in\Sigma^n} f_c \log f_c$$
$$-\sum_{xy\in\Sigma^2} f_{xy}^{ab} \log f_{xy}^{ab} + \sup_{d'\geq d+\epsilon} \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^{d'}.$$

The term $\sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^{d'}$ is the normalized log-likelihood of $\mathbb{A}[\{a,b\},M]$ given $d_T(a,b)=d'$. It can be easily verified that, under the Jukes–Cantor model, this is a concave function of d', with the maximum at $t'=-\frac{3}{4}\log(1-\frac{4}{3}f_{x\neq y})$ where $f_{x\neq y}=\sum_{xy\in\Sigma^2:x\neq y} f_{xy}^{ab}$ is the fraction of sites that differ between a and b (Felsenstein 2004). Because $f_{x\neq y}$ tends to $\frac{3}{4}-\frac{3}{4}e^{-4d/3}$ with probability 1, t' will tend to d. From the concavity of $\sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^{d'}$, we get

$$\sup_{d'\geq d+\epsilon} \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^{d'} = \sum_{xy\in\Sigma^2} f_{xy}^{ab} \log p_{xy}^{d+\epsilon}$$

with probability 1 as $m\to\infty$. This, together with the observation that $\sum_{c\in\Sigma^n} f_c \log f_c \to \sum_{c\in\Sigma^n} f_c^* \log f_c^*$ as $m\to\infty$, proves Inequality 3.

The proof of Inequality 4 proceeds analogously. ■

Lemma 3 provides us with the means to prove Lemma 1 and hence Theorem 1:

Proof of Lemma 1. By repeatedly applying Lemma 3 for all pairs of taxa, we obtain

$$\lim_{m\to\infty} \sup_{T:D(T,T^*)>\epsilon} \mathcal{L}(\mathbb{A}^m|T) \leq \sum_{c\in\Sigma^n} f_c^* \log f_c^*$$
$$-\min_{a,b\in N, s=\pm\epsilon} D_{KL}\left(\left\{f_{xy}^{ab}\right\} \middle|\middle| \left\{p_{xy}^{d+s}\right\}\right)$$

with probability 1. On the other hand, we have $\lim_{m\to\infty} \mathcal{L}(\mathbb{A}^m|T^*) = \sum_{c\in\Sigma^n} f_c^* \log f_c^*$ with probability 1, which proves Lemma 1.

Proof of Theorem 1. The proof proceeds analogously to the proof of Theorem 2 in the previous section. By repeatedly applying Lemma 1, we get

$$\Pr[D(ML(\mathbb{A}),T^*)<\epsilon]\to 1$$

for all $\epsilon>0$ which implies $D(ML(\mathbb{A}),T^*)\to 0$ as $m\to\infty$, as desired. ■

Conclusions

We have shown that standard maximum likelihood phylogenetic inference is statistically consistent even in the case where gaps are treated as missing data, as long as substitution rates across edges are nonzero. The problematic scenario highlighted by Warnow is not representative of the vast majority of data sets where substitutions are common. Given the widespread use of maximum likelihood in phylogenetic inference, this result should be reassuring to practitioners. Although it has been observed that alignment gaps can be used to reconstruct accurate phylogenies (Thatte 2006; Dessimoz and Gil 2010), our result shows that neglecting the information in gaps should not lead to incorrect inferences when sufficiently long sequences are available.

The methods used for the proof for gapped alignments have also permitted the construction of a new proof for ungapped alignments, which we consider simpler than other existing proofs.

Our proofs of consistency could be generalized to more complex evolutionary models. More flexible models of evolution, such as the general time-reversible model (Tavaré 1986) require inferring model parameters as well as the tree. To adapt our proof of Theorem 1 to this setting, one would have to define a suitable distance measure on the joint space of trees and substitution rate matrices. We leave this for future work.

In most practical scenarios, MSAs contain errors. Indeed, there are reasons to believe that this problem is more serious for large alignments, and manual curation requires a prohibitive amount of effort. Investigating whether MSAs introduce systematic biases in the long sequence limit is an interesting question for future research.

Funding

J.T. was supported equally by the Cancer Research UK Cambridge Institute and the European Molecular Biology Laboratory. N.G. was supported by the European Molecular Biology Laboratory.

Acknowledgments

We would like to thank Simon Tavaré for discussions. We would also like to thank Associate Editor Dr. Edward Susko and the anonymous reviewers whose suggestions helped improve the article.

References

Allison P.D. 2001. Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Sage Publications, Thousand Oaks, CA, United States.

Bouchard-Côté A., Jordan M.I. 2013. Evolutionary inference via the Poisson indel process. Proc. Natl Acad. Sci. USA 110:1160–1166.

Chang J.T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137:51–73.

Dessimoz C., Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.

Farris J.S. 1999. Likelihood and inconsistency. Cladistics 15:199–204.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Biol. 22:240–249.

Felsenstein, J. 2004. Inferring phylogenies. Sunderland: Sinauer Associates.

Grievink L.S., Penny D., Holland B.R. 2013. Missing data and influential sites: choice of sites for phylogenetic analysis can be as important as taxon sampling and model choice. Genome Biol. Evol. 5:681–687.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

MacKay D.J. 2003. Information theory, inference and learning algorithms. Cambridge University Press: Cambridge, United Kingdom.

McTavish E J., Steel M., Holder M.T. 2015. Twisted trees and inconsistency of tree estimation when gaps are treated as missing data—the impact of model mis-specification in distance corrections. Mol. Phylogenet. Evol. 93:289–295.

Miklós I., Lunter G., Holmes I. 2004. A long indel model for evolutionary sequence alignment. Mol. Biol. Evol. 21:529–540.

Rogers J.S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. 46: 354–357.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

RoyChoudhury A. 2014. Consistency of the maximum likelihood estimator of evolutionary tree. arXiv preprint arXiv:1405.0760.

Swofford D.L., Waddell P.J., Huelsenbeck J.P., Foster P.G., Lewis P.O., Rogers J.S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. In: Miura R.M., editor. Lectures on Mathematics in the Life Sciences. Providence: American Mathematical Society. p. 57–86.

Thatte B.D. 2006. Invertibility of the TKF model of sequence evolution. Math. Biosci. 200:58–75.

Thorne J.L., Kishino H., Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.

Wald A. 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20:595–601.

Warnow T. 2012. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. PLoS Curr. doi:10.1371/currents.RRN1308.

Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.