Phylogenetic analyses of alignments with gaps

Steve Evans[1] and Tandy Warnow*[2]

[1]Department of Statistics, University of California at Berkeley, Berkeley, CA, USA
[2]Department of Computer Science, University of Texas at Austin, Austin, Texas, USA

Email: Steven N. Evans - evans@stat.berkeley.edu; Tandy Warnow*- tandy@cs.utexas.edu;

*Corresponding author

Abstract

Background

Most statistical methods for phylogenetic estimation in use today treat a gap (generally representing an insertion or deletion, i.e., indel) within the input sequence alignment as missing data. However, the statistical properties of this treatment of indels has not been fully investigated.

Results

We prove that treating indels as missing data can be inconsistent for a general (and rather simple) model of sequence evolution, even when given the true alignment. We also prove that the true tree can be identified solely from the pattern of gaps in the true alignment (that is, character states can be ignored).

Conclusions

Our results show that the standard statistical techniques used to estimate phylogenies from sequence alignments may have unfavorable statistical properties, even when the sequence alignment is accurate and the assumed substitution model matches the generation model. Moreover, the pattern of gaps in an accurate alignment may give substantial information about the underlying

phylogeny, over and above what is present in the character states. These observations suggest that the recent focus on developing statistical methods that treat indel events properly is an important direction for phylogeny estimation.

Key words

Alignment, phylogeny estimation, indels, statistics.

Background

Phylogenetic estimation methods that analyze sets of molecular sequences generally have two steps: first, a multiple alignment of the sequences is estimated, and then a tree is estimated from the multiple alignment. When sequences have evolved under identifiable models, such as the General Time Reversible (GTR) model, then accurate estimations of trees with high probability is guaranteed, provided that appropriate methods (such as maximum likelihood) are used and the sequences are "long enough". Indeed, in practice, many nucleotide sequence phylogenies are estimated using statistical methods (maximum likelihood or Bayesian methods) that are based upon statistical models, such as GTR, for which these guarantees have been established.

Because biological evolution includes processes such as insertions and deletions (jointly referred to as "indels"), sequence datasets used in phylogenetic studies typically include sequences of different lengths. Therefore, phylogenetic analyses of sequence datasets first produce a multiple sequence alignment of the sequences via the addition of gaps, noted by '-', in the alignment. Once the sequences are aligned, a decision must be made about how to treat the gaps in the alignment.

In current practice, the following are the dominant gap-treatments:

- Remove all sites in which any gap appears, thus reducing to a gap-free alignment with fewer sites
- Assign an additional "fictitious" state for each gap.

- Code all the gaps in the alignment, and treat the presence/absence of gaps as a binary character (complementing the original sequence alignment character data).

- Treat the gaps as missing data. In parsimony analyses, this is often treated by finding the best nucleotide to replace the gap, but in likelihood-based analyses, this is often treated by summing the likelihood over all possible nucleotides for each gap.

The first option of removing all sites with gaps has the advantage of being statistically consistent for models in which the substitution process and the mechanism producing insertions and deletions are independent, but it has the disadvantage of removing data – and could result in sequence alignments that have so few sites as to be phylogenetically uninformative. Indeed, while this may not happen on small datasets, on large nucleotide datasets (especially if not for protein-coding markers), this could lead to empty alignments.

The second option of assigning an additional state for each gap presents other challenges. By definition, the true alignment represents positional homology, and hence two positions that have a nucleotide in a site constrain all nodes on the evolutionary path between them to also have a nucleotide in that position. In other words, ensuring that the model makes phylogenetic sense is rather complicated. Therefore, the substitution process must be extended carefully to handle an additional fictitious state properly. Finally, when the indel process can insert and delete several nucleotides in a row, the sites within the alignment no longer evolve independently, making this treatment invalid.

The third option, of coding each gap (maximal contiguous collection of dashes) in the alignment, includes a collection of techniques, ranging from extremely simple (create a single binary presence/absence character for each position that contains any gap) to very complex techniques. Software to automatically produce these additional binary characters encoding the gaps in a given alignment includes GapCoder [1], developed by Young and Healy, and also software developed for complex indel coding by Muller [2]. Simulation studies have shown improvements in tree estimation obtained through gap-coding over treating gaps as missing data (e.g., [3, 4]). However, the use of gap-coding is controversial (see the discussion in [3]), and not the dominant technique in phylogenetic analyses.

Instead, the most frequently used option (and the default for most software) is to treat gaps as missing data. Because of this, we focus our discussion on the impact of treating gaps

as missing data in phylogenetic analyses based upon Maximum Likelihood.

Treating Gaps as Missing Data

In this section, we discuss the statistical properties of estimating phylogenies using maximum likelihood, and treating gaps as missing data.

We begin with the Jukes-Cantor (JC) model of DNA sequence evolution model [5]. The JC model of site evolution assumes that only substitutions occur, and is characterized by a pair of parameters (T, θ), where T is a rooted binary tree with leaves labeled by a set S of taxa, and θ is a set of edge substitution probabilities, p_e (one for every edge $e \in E(T)$). Each substitution probability is constrained to satisfy $0 < p_e < \frac{3}{4}$; it gives the probability that the site changes on the edge e. The nucleotide at the root of the tree is selected from the uniform distribution over $\{A, C, T, G\}$. If the site changes its nucleotide state on edge e, then it changes with equal probability to one of the remaining three states. To use JC for modeling sequences, we assume that all sites evolve independently and identically ($i.i.d$).

Note, therefore, that JC model does not incorporate any mechanism for the formation of indels, so that sequences that are generated by this model will never have gaps.

Letting the tree topology T and alignment A be fixed, we define

$$ML_{JC}(\mathcal{A}, T) := \sup_{\theta} \mathbb{P}(\mathcal{A}|(T, \theta)).$$

That is, $ML_{JC}(\mathcal{A}, T)$ is the supremum of all likelihood scores obtained for JC model trees with the same fixed tree topology T (but allowing θ to vary). Although the likelihood is continuous the supremum may not actually be achieved for some θ because the range of values allowed for this parameter is not a closed set; that is, the supremum may be approached by parameter values θ for which some of the p_e are arbitrarily close to the boundary values 0 or $\frac{3}{4}$. Finally, we can talk about the JC maximum likelihood tree for a fixed gap-free alignment A, as the tree T such that the likelihood $ML_{JC}(\mathcal{A}, T)$ is maximized over all trees (and similarly we can define the EJC maximum likelihood tree).

Maximum likelihood inference of the parameter T under the JC model proceeds as follows:

- Input: sequence alignment $\mathcal{A}$ containing no gaps
- Output: all model trees T such that $ML_{JC}(\mathcal{A}, T)$ is maximized.

Let S be a set of DNA sequences in an alignment $\mathcal{A}$. We will say that the alignment A is *monotypic* if for each site in $\mathcal{A}$, there is exactly one nucleotide type (that is, all A's, all C's, all T's, or all G's). In particular, we do not allow any site to be entirely gapped. For example, the following is a monotypic alignment:

$$\begin{aligned} s_1 &= A-- \\ s_2 &= -C- \\ s_3 &= A-- \\ s_4 &= --- \\ s_5 &= --- \\ s_6 &= --T \\ s_7 &= A-- \end{aligned}$$

The following results were established in [6].

Lemma 1. *Let $\mathcal{A}$ be a monotypic alignment for the set S of sequences, and let T be an arbitrary tree on S. If gaps are treated as missing data, then $ML_{JC}(\mathcal{A}, T) = \left(\frac{1}{4}\right)^R$, where R is the number of sites in the alignment $\mathcal{A}$.*

Proof. This result follows from Lemma 1 in [6], but we sketch the proof here. For any tree T, the optimal settings of the edge substitution parameters on T have $p_e = 0$ for all edges (more correctly, the supremum $ML_{JC}(\mathcal{A}, T)$ is realized by a sequence of parameter values in which all the p_e converge to 0). For this setting of the substitution parameters, the probability of the data is just the probability of picking the correct state for that site, which is 1/4 under the JC model. Hence, the maximum likelihood score of the alignment, given the tree T, is $(\frac{1}{4})^R$, where R is the number of sites in $\mathcal{A}$. □

Theorem 1. *Let $\mathcal{A}$ be a monotypic alignment for set S. Then all trees on S are optimal solutions for maximum likelihood under JC, if gaps are treated as missing data.*

Proof. This result follows from Theorem 2 in [6], but we sketch the proof here. By Lemma 1, for monotypic alignments $\mathcal{A}$, the JC maximum likelihood scores for any tree are the same, so all trees are optimal solutions for maximum likelihood under JC. □

This theorem indicates a potential problem with treating gaps as missing data. If the mechanism generating the data has a high probability of producing aligned sequences that are monotypic or nearly so for some parameter values, then it will be difficult to reliably infer the underlying phylogenetic tree if the gaps are treated merely as missing data rather than features of the data that are informative about the path that evolution has taken. We address the issue of using the information present in the pattern of gaps more appropriately in the next section.

Estimating Trees using Indel Information

We now address the question of whether it is possible to estimate the true tree from the true alignment, using *just* the indel information. That is, we don't distinguish between nucleotides and only take into account whether it is a nucleotide or a gap that appears at a position in the alignment.

We begin by describing the continuous-time Markov chain model for the evolution of alignments used by Daskalakis and Roch [7]. At any time, the state of the process is a collection of sequences of nucleotides of equal length, with one sequence for each edge present in the underlying ultrametric tree at that time. For a given edge and site pair, the value of the sequence is either one of the nucleotides $\{A, C, T, G\}$ or the gap character $-$.

Each edge and site pair that has a nucleotide present gives birth independently at rate λ. When a site on an edge gives birth, it produces a copy of itself that is placed to the left or right with equal probability in the sequence of the individual alive on that edge. This produces a corresponding gap in the sequences of contemporaneous individuals on other edges. The result of a birth is a collection of sequences for each edge that is one site longer, for the sequence associated with the edge and site pair that gave birth the new site contains the nucleotide present in the old site, whereas the corresponding new site for all other edges contains the gap character $-$.

Similarly, each edge and site pair that has a nucleotide present dies independently at rate μ. The nucleotide for that edge and site pair is replaced by the gap character $-$ while all the states at that site for all other edges is left unchanged; this is unless replacing the nucleotide with a gap would produce a collection of sequences for which every edge has a

gap at the site, in which case the site is removed from all sequences and the common length of the sequences decreases by one.

Lastly, at bifurcations in the tree where a parent edge splits into two daughter edges, the sequence present at the end of the parent edge is copied to the start of each daughter edge, so that the number of sequences in the collection increases by one at such times.

The net result is that if the process starts with a single sequence drawn from the alphabet $\{A, C, T, G, -\}$ at the root, then it will produce a collection of equal length sequences, with one sequence for each each leaf (that is, taxon). It is possible that the collection of sequences produced in this way have zero length.

This is essentially the TKF91 birth-and-death model [8] except that there is no *immortal link* preventing the sequence length from being absorbed at zero. In particular, the process does not have an equilibrium distribution, and so there is no assumption that we are in equilibrium, as is the case in [8].

It is possible to combine the dynamics above with a substitution Markov chain that independently modifies the state of nucleotides along edges, but this will not be relevant for our purposes.

We can represent the state of the model at any time as a matrix with its entries drawn from the alphabet $\{A, C, T, G, -\}$, its rows indexed by the number of edges extant at that time, and its columns indexed by sites. The number of rows increases by one at any time that the tree T bifurcates, whereas the number of columns increases by one when a new site is born on some edge or decreases by one when the column for a a site would become filled with gap characters and hence is expunged.

We now make some simple observations about this model that will lead us to a statistical procedure for estimating the tree T.

Suppose first that the ultrametric tree just consists of a root a connected to two leaves b and c by edges with common length length ℓ. Assume that we start with a single site at the root and that there is a nucleotide at that site. The matrix seen at the leaves will have columns (which we will write for convenience as rows) of the type (N, N), $(N, -)$, and $(-, N)$, where N denotes that a nucleotide is present and $-$ denotes a gap. There is at most one site of the type (N, N). Such a site will be present if, after the bifurcation at the root

produces two copies of the nucleotide present at the root, this nucleotide is not removed by death events on either edge. The probability of such a site appearing at the leaves is thus $e^{-2\mu\ell}$.

Suppose now that we have an arbitrary rooted binary tree with root r. Let b and c be two leaves that are at common distance ℓ from their most recent common ancestor a. Write h for the distance from r to a. Assume as before that we start with a single site at the root and that there is a nucleotide at that site. The number of sites in the sequence at a that have a nucleotide present is the result of running for time h a birth-and-death process that increases by one (resp. decreases by one) at rate λk (resp. μk) when it is in state k. (The number of columns in the matrix keeping track of the the sequences at each point in the tree extant at a given time evolves in a considerably more complicated way, but if we just focus on the number of nucleotide (that is, non-gap) entries in the row corresponding to the ancestor of a at a given time, then this process has the stated dynamics.) In particular, the expected value of the number of sites in the sequence at a that have a nucleotide present is $\exp((\lambda-\mu)h)$. Therefore, the number of sites of type (N, N) (that is, that have a nucleotide present in the entries for both leaves b and c) has expected value $\exp((\lambda-\mu)h)\exp(-2\mu\ell)$.

By the same argument, the expected number of sites seen at the leaves b and c of type either (N, N) or $(N, -)$ (that is, that have a nucleotide present in the entry for the leaf b with the entry for the leaf c arbitrary) is $\exp((\lambda-\mu)(h+\ell))$. Similarly, the expected number of sites seen at the leaves b and c of type either (N, N) or $(-, N)$ (that is, that have a nucleotide present in the entry for the leaf c with the entry for the leaf b arbitrary) is also $\exp((\lambda-\mu)(h+\ell))$.

Suppose further that instead of starting with a single site at the root we now start with some number M of original sites in the sequence at the root (with nucleotides present at each site). Let p be the probability that the nucleotide present at some fixed site present at the root has a non-zero number of descendants at the leaves. Write X_{NN}, (resp. $X_{N\bullet}$ and $X_{\bullet N}$) for the number of columns in the matrix seen at the leaves that have nucleotides in the rows corresponding to the leaves b and c (resp. have a nucleotide in the row corresponding to b and have a nucleotide in the row corresponding to c). We have

$$\mathbb{E}[X_{NN}] = \frac{M}{p}\exp((\lambda-\mu)h - 2\mu\ell),$$

$$\mathbb{E}[X_{N\bullet}] = \frac{M}{p}\exp((\lambda-\mu)(h+\ell)),$$

and

$$\mathbb{E}[X_{\bullet N}] = \frac{M}{p}\exp((\lambda-\mu)(h+\ell)).$$

Write d for the distance on the tree T. When M is large we have, by the strong law of large numbers, that

$$\log\left(M^{-1}X_{NN}\right) \approx L_{NN} := (\lambda-\mu)d(r, b\wedge c) - \mu d(b,c) - \log(p),$$

$$\log\left(M^{-1}X_{N\bullet}\right) \approx L_{N\bullet} := (\lambda-\mu)d(r, b) - \log(p),$$

and

$$\log\left(M^{-1}X_{\bullet N}\right) \approx L_{\bullet N} := (\lambda-\mu)d(r, c) - \log(p),$$

where $b\wedge c$ is the most recent common ancestor of b and c.

Note that

$$\begin{aligned} L_{NN} &= \frac{1}{2}[(L_{N\bullet}+\log(p)) + (L_{\bullet N}+\log(p)) - (\lambda-\mu)d(b,c)] \\ &\quad - \mu d(b,c) - \log(p) \\ &= \frac{1}{2}[L_{N\bullet} + L_{\bullet N} - (\lambda+\mu)d(b,c)], \end{aligned}$$

and so

$$\begin{aligned} d(b,c) &= \frac{L_{N\bullet}+L_{\bullet N}-2L_{NN}}{\lambda+\mu} \\ &\approx \frac{1}{\lambda+\mu}\left(\log\left(M^{-1}X_{N\bullet}\right) + \log\left(M^{-1}X_{\bullet N}\right) - 2\log\left(M^{-1}X_{NN}\right)\right) \\ &= \frac{1}{\lambda+\mu}\log\left(\frac{X_{N\bullet}X_{\bullet N}}{X_{NN}^2}\right). \end{aligned}$$

Therefore, with enough data we can consistently estimate pairwise distances between taxa, and hence recover the tree T, from a knowledge of the true alignment without any assumptions about the substitution mechanism. This only requires a knowledge of the parameters λ and μ and does not require the values of M and p. Moreover, if we only want the tree up to a multiplicative scaling of distances (which is certainly enough to give the topology of the tree), then we don't even need to know λ and μ.

Discussion and Conclusion

The results in this paper show that treating gaps as missing data has the potential to result in meaningless phylogenetic estimations, since - under an extreme case in which the substitution probabilities are all zero - all trees are equally good solutions to maximum likelihood. On the other hand, even for such extreme cases, the indel process itself can contain information sufficient to identify the tree topology. Therefore, careful handling of sequence alignments that contain gaps is necessary.

What these results also show is that the current level of confidence in the systematics community regarding the desirable properties of statistically-based methods such as maximum likelihood and Bayesian MCMC need to be reconsidered when analyzing datasets that have evolved with indels. It seems likely that many of these analyses are quite reasonable, and the problematic results that are clearly possible in these cases may not apply (or at least not to the same extent) for these data. On the other hand, the current push within the phylogenetics research community to develop phylogenetic estimation methods that can co-estimate trees and alignments is encouraging, and could lead to improved statistical methods that can analyze large datasets, for estimating trees under appropriate models that include indels as well as substitutions, *given* accurate alignments.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SNE established results related to estimating trees from true alignments, using only the indel information, and TW established results related to the inconsistency of ML when treating gaps as missing data. TW and SNE wrote the paper.

Acknowledgements

This research was supported by two grants from the US NSF to TW (DEB 0733029 and DBI-1062335), the John Simon Guggenheim Memorial Foundation Fellowship to TW, a David Bruton Jr. Centennial Professorship to TW, and NSF grant (DMS 0907630) to SNE.

References

1. Young N, Healy J: **GapCoder automates the use of indel characters in phylogenetic analysis**. *BMC Bioinformatics* 2003, **4**(6).

2. Muller K: **Incorporating information from length-mutational events into phylogenetic analysis**. *Mol Phylogenet Evol* 2006, **38**:667–676.

3. Ogden TH, Rosenberg MS: **How should gaps be treated in parsimony? A comparison of approaches using simulation**. *Mol Phylo Evol* 2007, **42**:817–826.

4. Dwivedi B, Gadagkar S: **Phylogenetic inference under varying proportions of indel-induced alignment gaps**. *BMC Evol Biol* 2009, **9**:211.

5. Jukes TH, Cantor CR: **Evolution of protein molecules**. *Mammalian Protein Metabolism* 1969, :21–132.

6. Liu K, Warnow T, Holder M, Nelesen S, Yu J, Stamatakis A, Linder C: **SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees**. *Syst Biol* 2011. in press.

7. Daskalakis C, Roch S: **Alignment-free phylogenetic reconstruction**. In *RECOMB 2010, LNBI 6044*. Edited by Berger B, Springer-Verlag Berlin Heidelberg 2010:123–137.

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