Joint Bayesian Estimation of Alignment and Phylogeny

Benjamin D. Redelings, [1] **Marc A. Suchard** [2]

Keywords: Phylogeny, Alignment uncertainty, Bayesian, MCMC

We describe a model and algorithm for simultaneously estimating multiple alignments for biological sequences and the phylogenetic trees that relate the sequences. Unlike current techniques that base phylogeny estimates on a single best estimate of the alignment, we take into consideration the myriads of near-optimal alignments. We also avoid the trap of conditioning on an inaccurate external guide tree in constructing the alignment by estimating the alignment and phylogeny simultaneously. This eliminates the bias towards the guide tree that is inherent in phylogenies based on alignments constructed with progressive alignment [3]. The availability of the phylogeny during alignment construction also allows for more accurate models of both substitution and insertion/deletion that do not over-count single indels and substitutions that are shared between multiple taxa by common descent. Furthermore, this allows us to use shared indels as evidence in clustering taxa on the tree. We note that improved substitution models, such as those allowing invariant sites and rate variation between sites, may improve alignments in the joint estimation framework, whereas currently these models are only available in constructing phylogenies.

We use a continuous-time Markov chain process to describe the substitution process, with extensions for varying rates between sites. While current models implicitly condition on the alignment, we introduce an alignment prior, which allows us to treat the alignment as a parameter to be estimated. Our multiple alignments are built up from pairwise alignments along each branch of the tree. The alignment model is constructed from a hidden Markov model (HMM). We use a HMM with affine gap penalties, which avoids treating long indels as several unit-length indels. In addition to modeling alignments more accurately, this extension is important when using indels to group taxa as it does not exaggerate the number of rare events shared between taxa.

We take a Bayesian approach that allows us to estimate probable phylogenies and alignments, as well as measures of their support, by using Markov chain Monte Carlo (MCMC) techniques to sample from the joint posterior distribution for the phylogeny, alignment, and model parameters. We construct our Markov chain from straightforward Metropolis-Hastings steps for updating branch lengths and substitution parameters and several unique steps for updating the alignment and the topology that rely on dynamic programming. To update the alignment, we use modified versions of the two MCMC steps (branch alignment re-sampling, internal node re-sampling) proposed by [2]. In addition, we introduce a novel MCMC proposal to improve mixing that re-samples both a branch alignment and the internal node at one end of the branch. This proposal decreases burn-in substantially because it allows portions of the alignment to be aligned or unaligned without going through an unfavorable intermediate. We also introduce a new proposal to update the topology based on nearest-neighbor-interchange proposals, with some modifications to deal with internal nodes that lose definition when the topology is changed.

One problem that has intrigued molecular biologists is the question of whether the Archaea form a monophyletic group. To date, some analyses have supported monophyly of

[1]Department of Biomathematics, UCLA, Los Angeles, CA E-mail: `bredelin@ucla.edu`
[2]Department of Biomathematics, UCLA, Los Angeles, CA E-mail: `msuchard@ucla.edu`

Figure 1: Topology and alignment for a part of EF-Tu. Darker regions represent residues or gaps which are well resolved. Homo (a Eukaryote) and Sulfolobus (an eocyte) share an indel which is not present in the other Archaea, supporting paraphyletic Archaea.

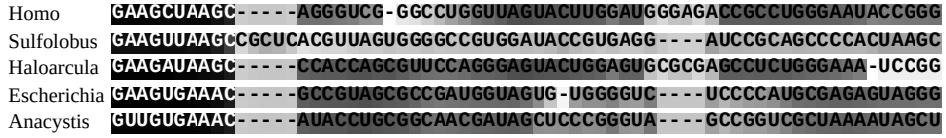

Figure 2: Alignment uncertainty for part of the 5S rRNA. Darker regions represent residues or gaps which are well resolved. The latter half of the alignment is ambiguous (shown), especially in regard to Sulfolobus. This makes the position of Sulfolobus difficult to resolve on the tree.

the Archaea, and some have placed the eocyte Archaea as sister taxa to Eukaryotes. This lack of resolution results partly from the fact that the inference depends on distantly related sequences that are difficult to align. Joint Bayesian estimation of alignment and phylogeny is an ideal method with which to approach this problem; joint estimation can deal with alignment ambiguity, avoids problems of bias in ambiguous alignments, and makes use of more information in the data than current phylogenetic reconstruction methods. To address the issue of the Archaea monophyly, we analyze both the 5S rRNA, and the EF-Tu/EF-1α gene. For each gene, we analyze data sets consisting of 5 taxa and 12 taxa.

The 5S rRNA left the location of the eocyte Archaea unresolved on the tree. However, based on EF-Tu/EF-1α, we find strong evidence against monophyly in that the eocytes are placed as sister taxa to the Eukaryotes (see Figure 1). Furthermore, we find strong support that the remaining Archaea are also paraphyletic. Our strong support for this topology stems from our methodology's use of evidence from common indels shared by eocytes and Eukaryotes. According to [1], about 75% of residues in EF-Tu have very well resolved homology; joint estimation can make use of the information from the 25% of residues with less resolved homology without being overconfident in their alignment. Our methodology can show alignment uncertainty in addition to uncertainty on trees. Figure 2 shows one important reason for the inability of the 5S rRNA to resolve the topology near the root; while some of the alignment is well resolved, approximately half of it is unusable for phylogenetic reconstruction because it is too ambiguous.

References

[1] Baldauf, S. L., Palmer, J. D., and Doolittle, W. F. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences USA* vol. 93, pp. 7749–7754

[2] Holmes, I. and Bruno, W. J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* vol. 17 no. 9 pp. 802–820

[3] Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution* vol. 8 pp. 378-385