BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–3

# HandAlign: Bayesian multiple sequence alignment, phylogeny, and ancestral reconstruction

Oscar Westesson, Lars Barquist*, Ian Holmes†

Department of Bioengineering, University of California Berkeley, CA, USA

Supplementary Version with Extended References

**ABSTRACT**

**Summary:** We describe `handalign`, a software package for Bayesian reconstruction of phylogenetic history. The underlying model of sequence evolution describes indels and substitutions. Alignments, trees, and model parameters are all treated as jointly-dependent random variables and sampled via Metropolis-Hastings Markov chain Monte Carlo (MCMC), enabling systematic statistical parameter inference and hypothesis testing. `handalign` implements several different MCMC proposal kernels, allows sampling from arbitrary target distributions via Hastings ratios, and uses standard file formats for trees, alignments and models.

**Availability and Implementation:** Installation and usage instructions are at `http://biowiki.org/HandAlign`

**Contact:** ihh@berkeley.edu

## 1 BACKGROUND

Multiple sequence alignments constitute a central part of many bioinformatics workflows. Commonly, an alignment is built from primary sequences, a tree is reconstructed from this alignment, and various analyses are run using the alignment and/or tree.

This can be problematic for several reasons. First, inference of the tree and alignment is likely to be *uncertain*: many alternative trees or alignments may explain the data comparably well. Downstream analyses which assume the tree and alignment to be true ignore this uncertainty, and potentially inherit embedded bias and error (Wong *et al.* (2008); Nelesen *et al.* (2008); Yang and dos Reis (2011); Malaspinas *et al.* (2011); Hanson-Smith *et al.* (2010)).

Second, this flow of information is *circular*: alignment algorithms often (implicitly or explicitly) make use of a guide tree, while tree- and model-fitting algorithms typically use an alignment as input. This leads to a chicken-and-egg situation (Varadarajan *et al.* (2008); Arribas-Gil (2010); Nelesen *et al.* (2008)).

In attempted resolution of these problems, the field of **statistical alignment** methods unifies alignment and tree-building as related inference tasks under a phylogenetic likelihood function (Hein (2001); Lunter *et al.* (2004, 2005); Fleissner *et al.* (2005)). The `handalign` software is one such tool, building on a range of prior work in this area (Holmes and Bruno (2001); Redelings and Suchard (2005); Bouchard-Côté *et al.* (2009)).

*Present address: Wellcome Trust Sanger Institute, Cambridge, UK

†To whom correspondence should be addressed

## 2 SAMPLING ALIGNMENTS, TREES, AND PARAMETERS

`handalign` implements a Bayesian model of sequence phylogeny with separate substitution and indel components. To perform inference of unknown variables (i.e. trees, ancestral sequences, or parameters) under this model, `handalign` makes use of Markov chain Monte Carlo sampling (MCMC). We cannot directly observe the indel history ($H$), tree ($T$), or evolutionary model parameters ($\theta$). However, we can estimate their *a posteriori* probability distribution, conditional on what we do observe: the extant sequences ($S$). That is, we aim to sample $(H, T, \theta|S)$. Explicitly marginalizing $H$, $T$ and $\theta$ is infeasibly expensive: there are combinatorially many trees $T$ and histories $H$, and continuously-varying parameters $\theta$. MCMC provides a powerful alternative way to sample from $(H, T, \theta|S)$ that is often not much more expensive than computing the joint likelihood of $(H, T, \theta, S)$.

Informally, MCMC randomly walks the space of $(H, T, \theta)$ tuples, the number of steps spent at a particular tuple converging to the posterior probability $P(H, T, \theta|S)$. The result is a series $\{(H_n, T_n, \theta_n)\}$ of samples from the posterior.

Depending on the investigator's goals, various analyses can then be performed using the collection of tuples $\{(H_n, T_n, \theta_n)\}$. The ensemble can be summarized with a single "consensus" alignment or tree,including confidence levels (e.g. the probability that a given subset of species form a monophyletic clade, or that a given column is correct) (Redelings and Suchard (2005)). Alternatively, downstream computations can be averaged over the ensemble: the sampled parameters can be used to estimate modes and moments (e.g. the most likely indel rate), or detect signatures of interest (e.g. positive selection).

As well as MCMC, `handalign` can perform a stochastic search using the same underlying model, but returning a single best-guess $(H, T, \theta)$ rather than a collection of such tuples.

## 3 CAPABILITIES

Given unaligned FASTA-format sequences, or (optionally) an "initial guess" in the form of a Stockholm-format alignment with an embedded Newick-format tree. `handalign` performs $N \times |S|$ sampling steps (where $|S|$ is the number of input sequences). Each step uses one of the MCMC kernel moves (described below) to update one of $H, T$, or $\theta$. If requested (via a command-line option), the new tuple $(H, T, \theta)$ is logged to a file. If operated in "Stochastic search" mode, a greedy local search is performed every $K$ samples to find the most likely nearby alignment. After $N \times |S|$ samples, the final $(H, T, \theta)$ tuple is output in Stockholm+Newick format.

© Oxford University Press 2011.

*Indel models:* The insertion-deletion model is an affine-gap transducer approximation (Holmes (2007)) to a Long Indel birth-death process (Miklós *et al.* (2004)) with insertion rate $\lambda$, deletion rate $\mu$, deletion extension probability $r$. The approximation is that indel events never overlap on the same branch. Other indel length distributions, such as TKF91 (Thorne *et al.* (1991)) or mixture-geometric, can be used.

*Substitution models:* Any parametric continuous-time Markov chain can be used to model character evolution, via the file format of the companion program `xrate` (Klosterman *et al.* (2006)). For instance, $20N$-state amino acid models (with $N$-valued hidden states (Holmes and Rubin (2002))) and 64-state codon models have been used. Parameters of these substitution models can be sampled, allowing alignment-free estimation of statistics such as $K_a/K_s$. Ancestral characters are summed out (Holmes and Bruno (2001)); they can be imputed using `xrate`.

*Tree prior:* The prior over tree topologies is uniform, with a weak exponential prior over total branch length. Alternate priors can be implemented using the "Arbitrary target distribution" mechanism.

*Arbitrary target distribution:* `handalign` allows any tree/alignment probability model to be implemented over a Unix pipe and used in a Metropolis-Hastings accept/reject step.

*MCMC kernel moves:* The relative proportions of the various sampling moves can be set on the command line. All are variants of Gibbs-sampling moves. Some are full Gibbs (perfectly mixing the sampled variables at every step); others utilize Metropolis-within-Gibbs (Gelfand (2000)) or a variant of importance sampling that includes the current point in the list of accessible points. The individual moves vary in the dependence of their complexities on the input sequence length, $L$. The worst-case complexity with default settings is $\mathcal{O}(L^2)$, comparable to BaliPhy (Redelings and Suchard (2005)).

*Stochastic search:* `handalign` can be used to do a partially-randomized greedy search, yielding a relatively quick, approximate maximum likelihood estimate for the alignment and/or tree, in addition to a full MCMC trace. The *iterative refinement* command-line option interrupts the MCMC run periodically to perform a greedy (Viterbi) search for the locally-maximal alignment close to the current sample.

*Alignment banding:* As the DP matrix may be costly to fill, both in time and memory, users may specify an alignment "band" as a heuristic constraint. Command-line options can be used to prevent visiting cells more than $M$ positions away from the current alignment path. This has the effect of causing indels longer than $M$ to be excluded, but is otherwise ergodic, and generally converts an $\mathcal{O}(L^a)$ step into an $\mathcal{O}(LM^{a-1})$ step (for $a \geq 2$).

*HMMoC adapter:* `handalign` can optionally make use of the Hidden Markov Model Compiler `hmmoc` (Lunter (2007)) to craft optimized C++ code for DP-based sampling steps. This typically speeds things up by a large constant factor.

*MCMC trace analysis:* The DART package includes several scripts for summarizing MCMC traces. `constock.pl` finds a consensus alignment and uses ANSI terminal color to render posterior probabilities of individual columns (Figure 1). Alternatively, trees can be extracted using `stocktree.pl` and a separate program, such as CONSENSE in the PHYLIP package, used to estimate consensus trees. `handalign` sampling traces use Stockholm alignment format to embed trees and parameters. A trace can be rendered as an ANSI terminal color animation using `stockfilm.pl`, converted into other common formats (see `http://biowiki.org/StockholmTools`) or the parameters extracted and their distribution analyzed (Figure 2).

*Current limitations and performance:* The $\mathcal{O}(L^2)$ complexity may be limiting for longer sequences (e.g. genomes); alignment banding should ameliorate this (but its effect on mixing performance is untested). Underflow/precision issues may potentially be an issue with larger trees. A ongoing compilation of comparison tests is here: `http://biowiki.org/HandAlignBenchmark`

## ACKNOWLEDGEMENT

The authors thank Benjamin Redelings, Marc Suchard, Jotun Hein, Joe Herman, Alexandre Bouchard-Côté and many others working in statistical alignment for their illuminating discussions.

*Funding*: Authors supported by NIH/NHGRI grant R01-GM076705.

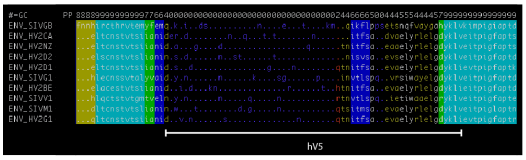


**Fig. 1.** A summary alignment of SIV/HIV gp120 proteins produced by `constock.pl`. Posterior probabilities of alignment columns are shown on the "PP" line (most significant decimal digit) and by ANSI terminal color (white-on-cyan is most reliable, blue-on-black is least). Hypervariable (hV) region 5 (Leonard *et al.* (1990)) corresponds with a low-confidence region.

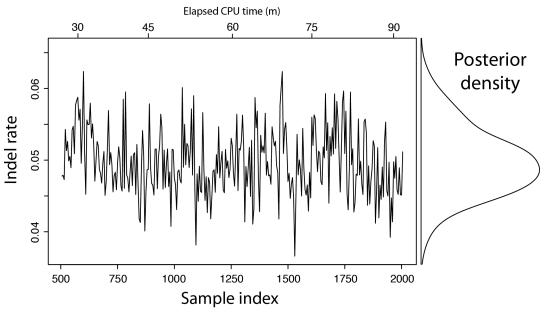


**Fig. 2.** The indel rate of the SIV/HIV gp120 protein has most of its probability mass concentrated between 0.04 and 0.06 indels per substitution. `handalign` was run for 90 minutes on a 3.4 GHz CPU, generating 2000 samples (500 discarded as burn-in). Every fifth sample is plotted; the entire trace was used to estimate the density.

## REFERENCES

Arribas-Gil, A. (2010). Parameter estimation in multiple-hidden IID models from biological multiple alignment. *Statistical Applications in Genetics and Molecular Biology*, **9**(1), 10.

Bouchard-Côté, A., Jordan, M. I., and Klein, D. (2009). Efficient inference in phylogenetic InDel trees. In *Advances in Neural Information Processing Systems 21 (NIPS)*, Vancouver, Canada.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005). Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst Biol*, **54**(4), 548–561. Comparative Study.

Gelfand, A. (2000). Gibbs sampling. *Journal of the American Statistical Association*, **95**(452), 1300–1304.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, **27**(9), 1988.

Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 179–190, Singapore. World Scientific.

Holmes, I. (2007). Phylocomposer and Phylodirector: Analysis and Visualization of Transducer Indel Models. *Bioinformatics*, **23**, 3263–3264.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**(9), 803–820.

Holmes, I. and Rubin, G. M. (2002). An Expectation Maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, **317**(5), 757–768.

Klosterman, P. S., Uzilov, A. V., Bendana, Y. R., Bradley, R. K., Chao, S., Kosiol, C., Goldman, N., and Holmes, I. (2006). XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, **7**(428).

Leonard, C., Spellman, M., Riddle, L., Harris, R., Thomas, J., and Gregory, T. (1990). Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in chinese hamster ovary cells. *Journal of Biological Chemistry*, **265**(18), 10373.

Lunter, G. (2007). HMMoC–a compiler for hidden Markov models. *Bioinformatics*, **23**(18), 2485–2487.

Lunter, G., Miklos, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83.

Lunter, G. A., Drummond, A. J., Miklós, I., and Hein, J. (2004). Statistical alignment: Recent progress, new applications, and challenges. In R. Nielsen, editor, *Statistical methods in Molecular Evolution*, Series in Statistics in Health and Medicine, chapter 14, pages 375–406. Springer Verlag.

Malaspinas, A., Eriksson, N., and Huggins, P. (2011). Parametric analysis of alignment and phylogenetic uncertainty. *Bulletin of mathematical biology*, **73**(4), 1–16.

Miklós, I., Lunter, G., and Holmes, I. (2004). A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, **21**(3), 529–540.

Nelesen, S., Liu, K., Zhao, D., Linder, C. R., and Warnow, T. (2008). The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. *Pacific Symposium on Biocomputing*, **2008**, 25–36.

Redelings, B. D. and Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, **54**(3), 401–418.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**, 114–124.

Varadarajan, A., Bradley, R. K., and Holmes, I. (2008). Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology*, **9**(10).

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, **319**(5862), 473–6.

Yang, Z. and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, **28**(3), 1217.